Politecnico di Milano

Facoltà di Ingegneria dei Sistemi
Corso di Laurea in Ingegneria Matematica

Tesi di Laurea Specialistica

Clustering di dati funzionali georeferenziati: studi di simulazione e analisi del database NASA "Earth Surface Meteorology and Solar Energy"

Candidato:
Alessia Pini
Matricola 735752

Relatore:
Dott. Simone Vantini

Correlatore:
Dott.ssa Valeria Vitelli

Anno Accademico 2009–2010

"As far as the laws of mathematics refer to reality, they are not certain;
and as far as they are certain, they do not refer to reality."
Albert Einstein

Ringraziamenti

Finalmente ci siamo! Ora che la mia carriera da studentessa del Politecnico volge al termine, vorrei ringraziare tutte le persone che mi hanno sostenuto in questi anni, grazie alle quali ho intrapreso e concluso questo bellissimo percorso, diventando quella che sono oggi.

Innanzitutto, un sentito ringraziamento va a tutti i professori del corso di studi di ingegneria matematica che ho potuto incontrare in questi anni, perché grazie alla loro disponibilità e passione nell'insegnamento mi hanno trasmesso anche un solo infinitesimo delle loro conoscenze, facendomi crescere dal punto di vista professionale e didattico, ma soprattutto umano.

Uno speciale ringraziamento va al prof. Piercesare Secchi, che mi ha dato la possibilità di lavorare su questa tesi, fidandosi di me il giorno in cui ho chiesto di poter lavorare con il suo gruppo. Grazie poi alla professoressa Anna Maria Paganoni, per la gentilezza e disponibilità che la contraddistinguono da sempre. Ringrazio infine di cuore Simone e Valeria, che mi hanno affidato questo progetto molto interessante, lasciandomi la possibilità di seguire il percorso che più mi interessava, e che mi hanno assiduamente seguito in questi mesi di lavoro, rispondendo tempestivamente ad ogni mio dubbio e dandomi sempre un prezioso aiuto e sostegno.

Ringrazio poi di cuore tutta la mia famiglia, per avermi sempre aiutato e supportato in ogni mia scelta. In particolare, grazie a mamma e papà, che hanno fatto di tutto per permettermi di sfruttare appieno ogni occasione. Grazie anche a mia sorella Ilaria, che ha in ogni occasione un modo per farmi tornare il sorriso: se non ci fosse dovrei inventarla. Infine, un grande abbraccio ai miei nonni, sui quali so che posso sempre contare.

Un grazie va anche a tutti gli amici che ancora mi sopportano nonostante i miei deliri matematici. Grazie a Laura, amica indispensabile, per esserci sempre, e per essere così com'è, a Lauretta, la mia fisio preferita, a Laura e Stefanino, con cui dai tempi di Mathonline ho condiviso esperienze splendide, alla Colzy (ehm... ok, non mi uccidere, volevo dire Valentina), a Marco e a Ferra, con la speranza di vivere insieme ancora nuove avventure.

Grazie a tutti gli amici che ho incontrato durante questi anni di Politecnico, Ecole Centrale e di nuovo Politecnico, che hanno reso le ore di lezione più piacevoli e gli esami da superare meno impossibili. Per citarvi tutti dovrei scrivere un'altra tesi... Comunque grazie di cuore! Spero (e in realtà ne sono convinta) che il legame che ho instaurato con molti di voi vada ben oltre questi 5 anni. Grazie, in particolare alla Dany, che continua ad avere l'altra metà del mio cervello (speriamo che almeno lei lo stia trattando bene), ad Ale, Cry (vi cito tutte e tre in un colpo solo!), Mary, Nao, Simone e Matteo, Flavietta, Silvia e Enrica (rigorosamente insieme), e chi più ne ha più ne metta. Pour conclure, merci à Tina et à tous les amis du 4G, qui ont rendu l'expérience centralienne inoubliable. J'espère de vour révoir bientot.

E dulcis in fundo...
Grazie a Luca. Per avermi aiutato a superare ogni problema durante la scrittura di questa tesi, per avermi iniziata all'utilizzo di BibTex, Dropbox, e altre amenità informatiche delle quali avrei altrimenti ignorato felicemente l'esistenza. Soprattutto grazie perché non so come farei senza di lui, e perché è la cosa più bella che mi sia capitata in questi ultimi anni.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

La statistica spaziale è quella branca della statistica che si occupa di analizzare, tramite l'introduzione di opportuni modelli, unità statistiche che sono associate a punti o aree nello spazio, come ad esempio la superficie terrestre. Si suppone quindi che i dati non siano tra loro indipendenti; al contrario, supponiamo che la dipendenza esistente tra i dati sia conseguenza della loro disposizione nello spazio: in questo senso, diremo che i dati sono spazialmente dipendenti. I dati vengono quindi analizzati tenendo opportunamente conto della posizione alla quale sono associati nello spazio preso in considerazione. Nel caso in cui tale posizione rappresenti un luogo sulla superficie terrestre indicato da coordinate geografiche, l'analisi condotta si colloca nell'ambito della geostatistica.

In questo lavoro di tesi si propone l'applicazione di nuove tecniche di statistica spaziale. Il lavoro, in particolare, si concentra su un problema di classificazione non supervisionata in ambito spaziale, in cui le unità statistiche sono di tipo funzionale. Più precisamente, supponiamo che la componente spaziale del dato si esprima attraverso un reticolo piano, che ricopre in modo continuo l'area di interesse (ad esempio un'area geografica), e ad ogni punto (o *sito*) del quale è associata un'osservazione. Ad ogni punto del reticolo si assume possa essere associata anche un'etichetta, che rappresenta alcune caratteristiche di interesse dell'area, dipendenti dal problema in esame: per esempio caratteristiche geografiche (l'area climatica), caratteristiche morfologiche del terreno, oppure la presenza o meno nel sottosuolo di una particolare sostanza chimica. Supponiamo infine di non essere a conoscenza delle etichette associate ad ogni punto del reticolo, che si dicono quindi nascoste, o latenti. Dalla ricostruzione di tali etichette si origina il problema di classificazione.

Per poter ricostruire le etichette latenti, supponiamo che sia possibile osservare l'evoluzione di un segnale osservato in corrispondenza di ogni sito, e legato alle caratteristiche di interesse; seguendo l'esempio dell'area geografica il segnale osservato potrebbe essere la temperatura, la pressione atmosferica, la piovosità o altre caratteristiche climatiche rispetto al tempo, oppure una risposta chimica nel dominio delle frequenze. L'ipotesi di lavoro è quindi che il segnale osservato in corrispondenza di ogni sito sia legato alla struttura spaziale sottostante esclusivamente attraverso l'etichetta associata allo stesso sito.

L'obiettivo dell'analisi è dunque ricostruire le etichette latenti esclusivamente tramite informazioni che derivano dal segnale funzionale osservato in ogni punto del reticolo. Se si suppone, come faremo nel seguito del lavoro di tesi, che le etichette appartengano ad un insieme discreto di cardinalità finita, per ricostruire le etichette latenti associate ad ogni sito in base al segnale osservato occorre classificare i dati di partenza in cluster costituiti da dati tra loro omogenei; la classificazione del segnale si riflette poi in una ricostruzione delle etichette del campo latente.

Per formalizzare la struttura spaziale che abbiamo brevemente descritto, introdurremo dei modelli statistici utilizzati in letteratura per descrivere la dipendenza spaziale

(si veda [Banerjee et al., 2004] e [Cressie, 1991]). Innanzi tutto ci porremo il problema di modellizzare la correlazione tra le osservazioni in siti differenti, che supponiamo essere una opportuna funzione della distanza tra due siti. Tale modello si basa sulla supposizione, abbastanza naturale, che la dipendenza tra i dati in due siti del reticolo sia tanto forte quanto più essi sono vicini e che, al contrario, diminuisca fino a tendere a zero all'aumentare della loro distanza. Da questa prima supposizione, si costruiscono dei modelli di dipendenza descritti attraverso un *covariogramma*, cioè una funzione che indica l'andamento della dipendenza spaziale al variare della distanza tra siti.

Un secondo modello che utilizzeremo per descrivere la dipendenza spaziale sono gli Hidden Random Markov Fields (HRMF), che riassumono concetti propri dei Markov Random Fields (MRF) da un lato, e degli Hidden Markov Models (HMM) dall'altro. In particolare, un Markov Random Field è una generalizzazione in due dimensioni del concetto di catena di Markov: si tratta di un modello in base al quale la dipendenza spaziale fra le variabili aleatorie agisce attraverso la definizione di un sistema di vicinanza (proprietà di Markov in due dimensioni). Nell'ambito del problema di ricostruzione del campo latente che abbiamo posto, possiamo quindi descrivere la dipendenza spaziale esistente tra le etichette tramite un unico modello MRF, che tuttavia non è osservabile, ma latente (o *hidden*). Da qui il parallelo con gli HMM, in cui i segnali osservati sono indipendenti condizionatamente agli stati dei siti cui sono associati, una volta nota cioè la realizzazione del campo Markoviano latente.

In tutto il lavoro di tesi supporremo, quindi, che i dati funzionali osservati siano realizzazioni di uno dei modelli precedentemente descritti. La tecnica di classificazione, proposta in [Secchi et al., 2011], che sarà utilizzata per ricostruire il campo latente delle etichette si basa su un'idea molto semplice ma efficace: invece di classificare i dati stessi, suddividiamo i siti in aree connesse, o tasselli, tramite la generazione di una *tassellazione di Voronoi*, ovvero utilizzando uno strumento matematico che prevede di scegliere casualmente un numero di siti (o centri) all'interno del dataset e di associare ogni altro sito al centro più vicino. In questo modo viene presa in considerazione la struttura di dipendenza spaziale, sfruttando l'idea che l'informazione sull'etichetta di un particolare sito non sia contenuta esclusivamente nel dato associato al sito stesso, ma anche nei dati associati ai siti vicini. Associamo quindi ad ogni tassello un dato funzionale che riassuma tutti i dati associati ai siti appartenenti a quel tassello, ed effettuiamo la classificazione con tecniche classiche sui dati di sintesi dei tasselli così ottenuti. Le prestazioni di tale tecnica saranno analizzate nel dettaglio nel corso del lavoro di tesi; la tecnica di classificazione sarà poi utilizzata per l'analisi di alcuni dataset reali, facenti parte del database NASA *Earth Surface Meteorology and Solar Energy*.

Il problema della classificazione di dati generati da un Hidden Random Markov Field è ampiamente discusso in letteratura nell'ambito della ricostruzione di immagini (si veda per esempio [Besag, 1974], [Geman and Graffigne, 1986], [Dubes et al., 1990], [Geman and Geman, 1993], [Künsch et al., 1995]). Questo lavoro di tesi, tuttavia, rappresenta un primo tentativo di integrare i modelli propri della geostatistica con l'analisi di dati funzionali (FDA), una recente area di ricerca nel campo statistico che attualmente riscontra grande interesse applicativo (si veda per esempio [Ferraty and Vieu, 2006], [Ramsay and Silverman, 2002], [Ramsay and Silverman, 2005]). L'aspetto innovativo del lavoro risiede infatti nella generalizzazione del problema di classificazione di dati spazialmente referenziati al caso in cui il segnale osservato e generato dal campo latente non sia multivariato, ma una realizzazione infinito dimensionale di una funzione aleatoria.

L'assenza in letteratura di modelli di riferimento per dati funzionali si accompagna

ad un'ulteriore difficoltà: spesso, nelle applicazioni a dati reali, non si conosce il numero di gruppi nei quali i dati vanno classificati, e dunque la cardinalità dell'insieme delle etichette che va stimata a partire dai dati stessi.

L'apporto innovativo di questo lavoro di tesi si articola secondo due linee principali. Innanzi tutto, la validazione tramite le simulazioni effettuate nel Capitolo 3 dell'algoritmo proposto, nonché lo studio specifico del comportamento di tale tecnica di classificazione al variare dei parametri del modello utilizzato per generare i dati, e dei parametri dell'algoritmo stesso. Un secondo contributo fortemente innovativo del lavoro di tesi riguarda la declinazione dello schema generale presentato al caso reale trattato, nel quale i siti non sono disposti su di un reticolo piano, ma sulla superficie di una sfera che rappresenta la superficie terrestre. Per poter effettuare le analisi sui dataset reali considerati, infatti, è necessario adattare l'algoritmo proposto in modo da simulare una tassellazione di Voronoi sulla sfera, invece che sul piano, tenendo contestualmente conto del fatto che i siti nei quali si suddivide il reticolo sono relativi a superfici la cui dimensione varia al variare della latitudine.

Il lavoro di tesi è articolato nel modo seguente: nel **Capitolo 1** introduciamo i concetti necessari alla formalizzazione dei modelli citati in ambito spaziale. In particolare, forniamo una panoramica generale dei diversi modelli esistenti per il covariogramma, e introduciamo i modelli di Markov Random Fields notevoli, come i *campi di Ising*.

Nel **Capitolo 2**, introduciamo la tecnica esplorativa proposta in [Secchi et al., 2011], come uno strumento volto ad effettuare la classificazione di dati funzionali sfruttando opportunamente tutte le informazioni fornite dalla struttura spaziale esistente.

Nel **Capitolo 3**, analizziamo le prestazioni della tecnica di classificazione proposta nel Capitolo 2, tramite la generazione e l'analisi di dataset sintetici che presentino la dipendenza spaziale descritta dai modelli presentati nel Capitolo 1. In particolare, simuliamo le etichette latenti come realizzazioni di un campo di Ising; vengono poi generati i dati funzionali, indipendenti o meno condizionatamente alle etichette. Utilizziamo la tecnica proposta per la classificazione dei dati così generati; i risultati ottenuti vengono confrontati, in termini di rate di misclassificazione, con quelli ottenuti classificando i dati con tecniche classiche (che non utilizzano le informazioni provenienti dalla struttura spaziale esistente). In particolare, siamo interessati allo studio delle prestazioni della tecnica proposta al variare di alcune caratteristiche del modello di generazione dei dati, quali il modello di dipendenza spaziale, o la distribuzione di emissione del segnale. Inoltre, un altro obiettivo delle simulazioni effettuate è lo studio del comportamento dell'algoritmo al variare di alcuni parametri costitutivi dell'algoritmo stesso, quali il numero di tasselli utilizzati per la tassellazione di Voronoi.

Nel **Capitolo 4**, infine, utilizziamo la tecnica di classificazione proposta ed analizzata per l'analisi statistica di un caso reale. Ci proponiamo di analizzare alcuni dati ottenuti dal database NASA *Earth Surface Meteorology and Solar Energy*, che contiene osservazioni di alcuni parametri di riferimento per le applicazioni nell'ambito dell'energia solare, in corrispondenza di un reticolo di siti che ricopre l'intera superficie terrestre. In un primo momento, ci proponiamo di classificare un dataset composto da osservazioni riguardanti l'evoluzione nel tempo del numero di giorni senza sole consecutivi al mese, un parametro di grande importanza per il dimensionamento e la progettazione di impianti ad energia fotovoltaica. Successivamente analizziamo, con la stessa tecnica, un altro dataset costituito, a partire dal dataset precedente, da osservazioni riguardanti la quantità massima di energia che una batteria collegata ad un impianto fotovoltaico deve poter immagazzinare per il funzionamento a regime costante.

Capitolo 1

Modelli di dipendenza spaziale

1.1 Introduzione alla geostatistica

L'oggetto dell'analisi svolta nella tesi è la classificazione di dati con dipendenza spaziale. In particolare, quindi, nel corso di questo lavoro di tesi, tratteremo dati osservati in una particolare posizione spaziale, e l'obiettivo del lavoro sarà effettuare delle analisi statistiche che tengano conto della posizione nello spazio di riferimento nella quale i dati sono stati osservati. La branca della statistica che si occupa dell'analisi di tali dati prende il nome di statistica spaziale, o geostatistica nel caso in cui si trattino dati geografici.

Nel corso del primao capitolo, presentiamo quindi i principali modelli utilizzati in letteratura per descrivere la struttura di dipendenza esistente tra i dati stessi. Successivamente, nel Capitolo 3 utilizzeremo i modelli introdotti per creare dei dataset sintetici composti da dati spazialmente dipendenti, al fine di testare le prestazioni delle tecniche di classificazione che introdurremo.

Più precisamente, siamo interessati alla modellizzazione di dati multivariati o funzionali che abbiano particolari caratteristiche comuni:

- siano spazialmente referenziati, cioè associati ad informazioni riguardanti la posizione spaziale (che può essere puntuale o continua, associata ad una determinata regione) nella quale tali dati sono stati osservati. Nella maggior parte delle applicazioni la posizione spaziale alla quale sono associati i dati è una posizione geografica, in termini di latitudine e longitudine. Si parla allora di dati georeferenziati;

- la struttura di dipendenza dei dati tenga conto della posizione nello spazio. Un'ipotesi abbastanza verosimile, e che viene in generale assunta nella pratica, è che i dati non siano tra loro indipendenti, e che la correlazione tra dati osservati in posizioni vicine sia maggiore di quella tra dati osservati in posizioni lontane.

Analizzeremo in particolare due diversi tipi di modelli spaziali che vengono in genere utilizzati in ambiti leggermente diversi, ovvero modelli di dipendenza descritti attraverso un *variogramma* da un lato e *Hidden Random Markov Fields* dall'altro. In entrambi i casi, l'oggetto dell'inferenza statistica è un particolare processo stocastico $\{Y : \mathcal{D} \to \mathcal{E}\}$, dove $\mathcal{D}$ è un sottospazio di uno spazio metrico r-dimensionale e rappresenta l'area all'interno della quale osserviamo i dati ed $\mathcal{E}$ è un opportuno sottospazio di $\mathbb{R}^p$ o uno spazio funzionale. Supponiamo di osservare il processo Y in corrispondenza di N punti o *siti* dell'area $\mathcal{D}$, che chiameremo $\{\mathbf{s}_1, ..., \mathbf{s}_N\}$, dove $\mathbf{s}_i \in \mathcal{D}$ rappresenta la coordinata spaziale

dell'i–esimo dato. In generale, la disposizione dei siti all'interno di $\mathcal{D}$ può non seguire uno schema regolare. Se per esempio, l'analisi statistica riguarda la concentrazione di un inquinante atmosferico nelle maggiori città italiane, i siti $\mathbf{s}_i$ corrispondono alle coordinate geografiche delle città prese in considerazione, che non si dispongono ovviamente secondo uno schema fisso. Nei problemi che prenderemo in considerazione invece supporremo di avere a disposizione osservazioni registrate in corrispondenza dei punti di un reticolo regolare.

In generale, per l'analisi di dati geostatistici, $\mathcal{D}$ è un sottospazio di $\mathbb{R}^2$ dotato di distanza euclidea se si considerano aree geografiche di piccole dimensioni (per le quali l'effetto della curvatura terrestre è trascurabile); se invece si considerano aree geografiche molto grandi, $\mathcal{D}$ è una parte o l'intera superficie di una sfera dotata di distanza geodetica. A seconda delle applicazioni, in alternativa, risulterà più comodo utilizzare un altro sottospazio, o un'altra distanza.

I dati che tratteremo sono quindi una realizzazione del processo Y in un sottoinsieme discreto di punti di $\mathcal{D}$, $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\}$. I modelli che prenderemo in considerazione per $Y(\mathbf{s}_1), ..., Y(\mathbf{s}_N)$ hanno diverse proprietà a seconda delle ipotesi fatte sulla regolarità dell'insieme dei punti in cui il processo viene osservato.

1.2 Modelli di dipendenza descritti attraverso un variogramma

1.2.1 Stazionarietà e covariogramma

In un primo momento, non facciamo nessuna ipotesi sulla struttura del sottoinsieme $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\} \subset \mathcal{D}$, salvo l'esistenza di una distanza d su $\mathcal{D}$. Assumiamo che il processo osservato Y abbia valori in un sottoinsieme $\mathcal{E}$ di $\mathbb{R}$, cioè, $p = 1$. Assumiamo inoltre che Y abbia media $\mathbb{E}[Y(\mathbf{s})] = \mu(\mathbf{s})$. Diamo le seguenti definizioni rispettivamente di stazionarietà forte e stazionarietà debole (o di secondo ordine).

Definizione 1.1. *Il processo Y è strettamente stazionario se $\forall n \geq 1$, per ogni insieme di n siti $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ e ogni $\mathbf{h} \in \mathbb{R}^r$, i vettori aleatori $(Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$ e $(Y(\mathbf{s}_1 + \mathbf{h}), ..., Y(\mathbf{s}_n + \mathbf{h}))$ hanno la stessa distribuzione congiunta.*

Definizione 1.2. *Il processo Y è debolmente stazionario se $\mu(\mathbf{s}) \equiv \mu \quad \forall \mathbf{s} \in \mathcal{D}$ e $\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$, per tutti gli $\mathbf{h} \in \mathbb{R}^r$ e $\mathbf{s} \in \mathcal{D}$ tali che $\mathbf{s} + \mathbf{h} \in \mathcal{D}$.*

La stazionarietà debole garantisce quindi che la covarianza tra i valori del processo in ogni coppia di punti del dominio $\mathcal{D}$ possa essere riassunta dalla funzione $C(\mathbf{h})$, che dipende esclusivamente dal vettore di separazione tra i due punti e non dai punti stessi. Osserviamo inoltre che la proprietà di stazionarietà debole è meno restrittiva di quella forte, ed è implicata da quest'ultima. Nel caso in cui valga l'ipotesi di stazionarietà debole, la funzione di covarianza $C(\mathbf{h})$ è detta *covariogramma* del processo Y.

1.2.2 Stazionarietà intrinseca e variogramma

Diamo ora la definizione di un terzo ed ultimo tipo di stazionarietà del processo Y, che prende il nome di stazionarietà intrinseca.

Definizione 1.3. *Il processo Y è intrinsecamente stazionario se, per ogni $\mathbf{h} \in \mathbb{R}^r$ e $\mathbf{s} \in \mathcal{D}$ tali che $\mathbf{s} + \mathbf{h} \in \mathcal{D}$, $\mathbb{E}[Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})] = 0$ e inoltre si ha:*

$$\mathbb{E}[(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s}))^2] = \mathrm{Var}(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})) = 2\gamma(\mathbf{h}).$$

In breve, la definizione 1.3 implica che la varianza della differenza tra il valore del processo in due punti qualsiasi di $\mathcal{D}$ dipenda esclusivamente dal vettore di separazione tra i due punti. Se il processo Y è intrinsecamente stazionario, la funzione $2\gamma(\mathbf{h})$ è detta *variogramma* del processo (quindi $\gamma(\cdot)$ è detto *semivariogramma*).

Esploriamo meglio la relazione tra le ultime due definizioni di stazionarietà. Per prima cosa, ipotizziamo che valga 1.2, quindi che il processo Y sia debolmente stazionario ed esista il covariogramma $C(\mathbf{h})$. Allora:

$$\begin{aligned}\mathrm{Var}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) &= \mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s})) - 2\,\mathrm{Cov}(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})].\end{aligned}$$

La varianza di $Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})$ non dipende quindi dal punto $\mathbf{s}$ ma esclusivamente da $\mathbf{h}$. Ricaviamo dunque che, senza ulteriori ipotesi, il processo Y è anche intrinsecamente stazionario e il semivariogramma è dato da:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \tag{1.1}$$

Abbiamo mostrato che la stazionarietà debole implica quella intrinseca. Per vedere ora sotto quali ipotesi vale il viceversa, supponiamo che Y sia intrinsecamente stazionario. Possiamo ricavare un'espressione per C, conoscendo γ? In generale, sia:

$$\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) =: c(\mathbf{s}, \mathbf{h}). \tag{1.2}$$

Vogliamo mostrare, sotto quali ipotesi la funzione $c(\mathbf{s}, \mathbf{h})$ definita in 1.2 sia funzione del solo vettore $\mathbf{h}$. Sotto l'ipotesi di ergodicità del processo spaziale, cioè se supponiamo che la covarianza tra i valori del processo Y in due punti tenda a zero se la distanza tra i due punti diventa molto grande, abbiamo:

$$\lim_{\|\mathbf{h}\|\to\infty} \mathrm{Cov}(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) = \lim_{\|\mathbf{h}\|\to\infty} c(\mathbf{s}, \mathbf{h}) = 0. \tag{1.3}$$

Dall'ipotesi di stazionarietà intrinseca (Definizione 1.3), abbiamo poi, $\forall \mathbf{h} \in \mathbb{R}^r, \mathbf{s} \in \mathcal{D}$ tali che $\mathbf{s}+\mathbf{h} \in \mathcal{D}$:

$$\mathrm{Var}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) = 2\gamma(\mathbf{h}). \tag{1.4}$$

Abbiamo poi, in generale,

$$\begin{aligned}2\gamma(\mathbf{h}) &= \mathrm{Var}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) \\ &= \mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s})) - 2\,\mathrm{Cov}(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) \\ &= \mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s})) - 2c(\mathbf{s}, \mathbf{h}),\end{aligned}$$

da cui:

$$c(\mathbf{s}, \mathbf{h}) = \frac{1}{2}(\mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s}))) - \gamma(\mathbf{h}). \tag{1.5}$$

A questo punto, dall'ipotesi di ergodicità del processo Y (equazione 1.3), passando al limite per $\|\mathbf{h}\| \to \infty$ in 1.5, abbiamo $\lim_{\|\mathbf{h}\|\to\infty} c(\mathbf{s}, \mathbf{h}) = 0$, da cui:

$$\begin{aligned}2 \lim_{\|\mathbf{h}\|\to\infty} \gamma(\mathbf{h}) &= \lim_{\|\mathbf{h}\|\to\infty} \mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s})) \\ &= \lim_{\|\mathbf{h}\|\to\infty} c(\mathbf{s}+\mathbf{h}, \mathbf{0}) + c(\mathbf{s}, \mathbf{0}),\end{aligned}$$

dove, l'ultima uguaglianza deriva dalla definizione della funzione $c(\mathbf{s}+\mathbf{h}, \mathbf{0})$, nell'equazione 1.2. Dall'ipotesi di ergodicità (equazione 1.3), si ha poi $\lim_{\|\mathbf{h}\|\to\infty} c(\mathbf{s}+\mathbf{h}, \mathbf{0}) =$

0. In più, sappiamo dall'ipotesi di stazionarietà intrinseca che $\gamma(\mathbf{h})$ non dipende da $\mathbf{s}$, quindi se esiste finito il limite $\lim_{\|\mathbf{h}\|\to\infty}\gamma(\mathbf{h})$, non dipende da $\mathbf{s}$. In particolare, quindi, anche $c(\mathbf{s},\mathbf{0}) =: \tilde{C}(\mathbf{0})$ non dipende da $\mathbf{s}$. Dalle ipotesi fatte, abbiamo:

$$\mathrm{Var}(Y(\mathbf{s}+\mathbf{h})) + \mathrm{Var}(Y(\mathbf{s})) = \tilde{C}(\mathbf{0}) = 2\lim_{\|\mathbf{u}\|\to\infty}\gamma(\mathbf{u}),$$

da cui:

$$c(\mathbf{s},\mathbf{h}) = \lim_{\|\mathbf{u}\|\to\infty}\gamma(\mathbf{u}) - \gamma(\mathbf{h}) =: C(\mathbf{h}) \tag{1.6}$$

Dove, l'ultimo passaggio vale poiché il termine a destra dell'uguale non è funzione di $\mathbf{s}$. In generale, non è detto che il limite a destra esista. Se questo è il caso, allora il processo è anche debolmente stazionario e il covariogramma è dato da 1.6.

Le implicazioni tra i diversi tipi di stazionarietà sono riassunti in Figura 1.1.

Figura 1.1: Implicazioni tra stazionarietà forte, debole e intrinseca

1.2.3 Isotropia

Un ultimo concetto importante per la modellizzazione di dati spaziali è quello di *isotropia*.

Definizione 1.4. *Sia $Y : \mathcal{D} \to \mathbb{R}$ un processo intrinsecamente stazionario. Il processo si dice isotropo se il variogramma $\gamma(\mathbf{h})$ dipende dal vettore di separazione $\mathbf{h}$ solo attraverso il suo modulo $\|\mathbf{h}\|$: $\gamma(\mathbf{h}) = \gamma(\|\mathbf{h}\|)$.*

Ciò vuol dire che la dipendenza spaziale è funzione esclusivamente della distanza tra i punti, e non dipende, per esempio, dalla direzione. Per i processi isotropi, quindi, il semivariogramma (e se esiste, il covariogramma) può essere espresso in funzione della sola variabile $\|\mathbf{h}\|$, che per semplicità di notazione rinomineremo t.

Questa proprietà permette quindi di esprimere variogramma e covariogramma in maniera molto semplice, in funzione di alcuni parametri che caratterizzano la dipendenza spaziale: in letteratura sono state proposte alcune forme parametriche tra le quali è possibile scegliere un valido variogramma, o semivariogramma (si veda, per esempio [Banerjee et al., 2004]). Descriviamo ora brevemente quelle più importanti, mostrando il significato dei parametri utilizzati.

- Semivariogramma lineare:

$$\gamma(t) = \begin{cases} 0 & t = 0 \\ \tau^2 + \sigma^2 t & t > 0 \end{cases} \tag{1.7}$$

  Il semivariogramma lineare è la forma più semplice in assoluto. Tuttavia, osserviamo che $\lim_{t\to\infty}\gamma(t) = \infty$: questo è un esempio di processo intrinsecamente ma non debolmente stazionario. In questo caso non esiste la funzione C.

- Semivariogramma sferico:

$$\gamma(t) = \begin{cases} 0 & t = 0 \\ \tau^2 + \sigma^2 \left[\frac{3\Phi t}{2} - \frac{(\Phi t)^3}{2} \right] & 0 < t \leq 1/\Phi \\ \tau^2 + \sigma^2 & t \geq 1/\Phi \end{cases} \tag{1.8}$$

  Il semivariogramma sferico è valido in uno spazio a $r = 1, 2$ o 3 dimensioni. Per $r \geq 4$ genera una matrice di covarianza che non è definita positiva.

  Questo semivariogramma fornisce tuttavia un'illustrazione intuitiva della diversa funzione dei tre parametri specificati, cioè $\lambda = 1/\Phi$, σ e τ, rispettivamente *range*, *partial sill* e *nugget*. Per definizione, $\gamma(0) = 0$, ma la funzione γ è discontinua nell'origine, e tale discontinuità è quantificata dal nugget: $\lim_{t \to 0^+} \gamma(t) = \tau^2$. Il valore asintotico che raggiunge la funzione γ, invece, è espresso dal sill $\tau^2 + \sigma^2$ (mentre il solo σ^2 è detto partial sill). Il range, infine, rappresenta la distanza minima t alla quale due punti devono essere affinché $\gamma(t)$ raggiunga il livello massimo. Infatti per $t \geq \lambda$, $\gamma(t) = \tau^2 + \sigma^2$.

  Questo semivariogramma, a differenza di quello lineare, verifica l'ipotesi $\lim_{t \to \infty} < \infty$, quindi, sotto l'ipotesi di ergodicità di Y, possiamo calcolare il covariogramma $C(t)$:

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & t = 0 \\ \sigma^2 \left[1 - \frac{3\Phi t}{2} + \frac{(\Phi t)^3}{2} \right] & 0 < t \leq 1/\Phi \\ 0 & t \geq 1/\Phi \end{cases} \tag{1.9}$$

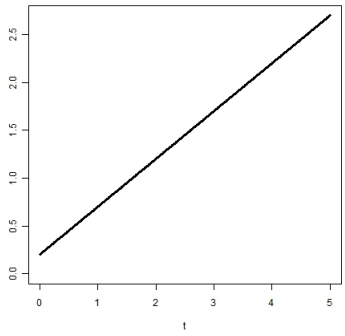
  Osservando l'espressione del covariogramma, riconsideriamo i concetti di sill, nugget e range: il sill è i valore della varianza puntuale in $t = 0$ e si scompone in partial sill, che entra nella matrice di covarianza moltiplicato per un fattore che decresce in t, e nugget τ^2, che è la parte di tale varianza che entra in gioco esclusivamente per $t = 0$. Per questo motivo, il nugget τ^2 può essere visto come la parte di varianza non spaziale, mentre il partial sill σ^2 come la parte di varianza spaziale. Il range $\lambda = 1/\Phi$ infine è il minimo valore di t a partire dal quale non c'è più correlazione tra due punti.

- Semivariogramma esponenziale:

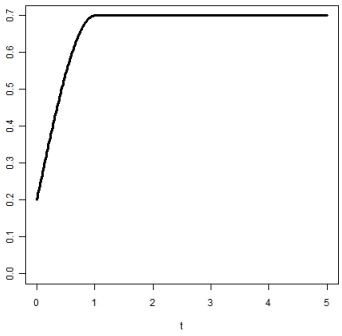
$$\gamma(t) = \begin{cases} 0 & t = 0 \\ \tau^2 + \sigma^2[1 - \exp(-\Phi t)] & t > 0 \end{cases} \tag{1.10}$$

  Il vantaggio del semivariogramma esponenziale su quello sferico è che l'espressione funzionale è più semplice, e genera matrici di covarianza valide in tutte le dimensioni. In questo caso, mentre sill e nugget hanno esattamente la stessa interpretazione di quella fornita per il semivariogramma sferico, c'è una leggera differenza per quanto riguarda il range λ. Infatti, il valore asintotico $\tau^2 + \sigma^2$ non è mai raggiunto, quindi il range come l'abbiamo definito precedentemente è infinito. In casi come questo, si utilizza invece la nozione di *range effettivo*, cioè la distanza t a partire dalla quale non c'è essenzialmente incremento in $\gamma(t)$ (quindi, passando al covariogramma, la distanza a partire dalla quale i valori nel processo in due punti non sono più correlati). Il range effettivo viene comunemente preso in corrispondenza del valore numerico $\frac{3}{\Phi}$, poiché il valore numerico della funzione $C(\cdot)$ a tale distanza è circa 0.05. Anche in questo caso, sotto l'ipotesi di ergodicità del processo Y, il covariogramma esiste, ed è dato da:

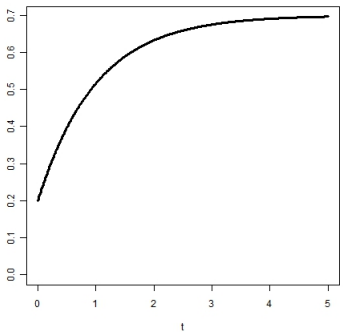
$$C(t) = \begin{cases} \tau^2 + \sigma^2 & t = 0 \\ \sigma^2 \exp(-\Phi t) & t > 0 \end{cases} \tag{1.11}$$



(a) Lineare



(b) Sferico

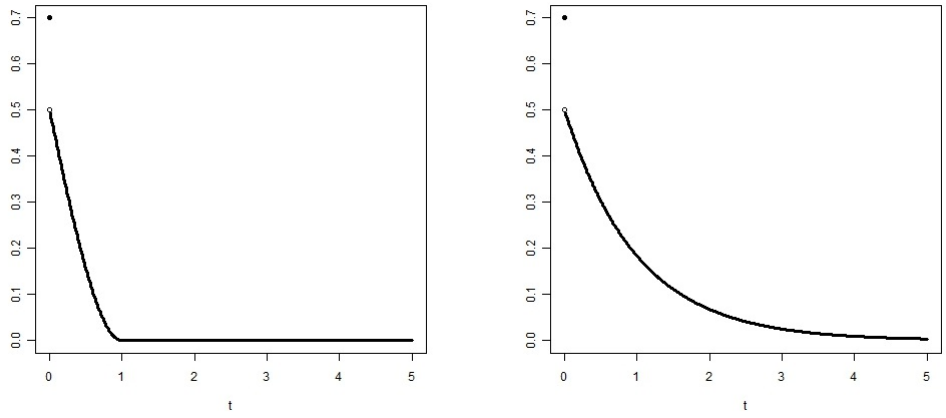


(c) Esponenziale

Figura 1.2: Semivariogramma lineare (a), sferico (b) ed esponenziale (c) per i parametri $\tau^2 = 0.2$, $\sigma^2 = 0.5$, $\Phi = 1$

In Figura 1.2 mostriamo i tre semivariogrammi descritti in funzione della distanza t. Confrontando i tre grafici, nei quali i parametri scelti sono gli stessi ($\tau^2 = 0.2$, $\sigma^2 = 0.5$, $\Phi = 1$), osserviamo che innanzi tutto il semivariogramma lineare diverge al crescere di t, mentre gli altri due semivariogrammi tendono verso un valore asintotico che è il sill, cioè $\sigma^2 + \tau^2$. Osserviamo inoltre che, menre il semivariogramma sferico raggiunge il valore asintotico per $t = 1 = \frac{1}{\Phi}$, il semivariogramma esponenziale tende solo asintoticamente a tale valore, e la crescita è più lenta. Osserviamo infine che il semivariogramma esponenziale raggiunge alori molto vicini al sill a partire da $t \simeq \frac{3}{\Phi} = 3$.

In Figura 1.3 presentiamo i covariogrammi, sempre con la scelta di parametri $\tau^2 = 0.2$, $\sigma^2 = 0.5$, $\Phi = 1$ nel caso sferico e esponenziale (ricordiamo che il covariogramma lineare non è definito poiché il variogramma è illimitato). Osserviamo che quanto avevamo detto commentando la funzione dei parametri di sill, nugget e range si riflette anche nella forma del covariogramma (il sill è il valore $C(0)$, il nugget la discontinuità nell'origine e il range la lunghezza a partire dalla quale il covariogramma è nullo nel caso sferico e ha valori molto piccoli nel caso esponenziale).



(a) Sferico (b) Esponenziale

Figura 1.3: Covariogramma sferico (a) ed esponenziale (b) per i parametri $\tau^2 = 0.2$, $\sigma^2 = 0.5$, $\Phi = 1$

1.3 Modelli spaziali tramite Hidden Random Markov Fields

1.3.1 Introduzione agli Hidden Random Markov Fields

Presentiamo ora un altro modello che utilizzeremo per la modellizzazione della dipendenza spaziale dei dati. In questo caso, la dipendenza spaziale, invece che essere introdotta direttamente specificando la covarianza tramite un variogramma o covariogramma, viene introdotta in maniera gerarchica.

Nel Paragrafo 1.3.2 presenteremo un modello per un processo spaziale $\{X(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ a valori in un insieme discreto $\mathcal{L}$, con una struttura di dipendenza spaziale tra siti vicini che definiremo. Successivamente forniremo un modello per il processo osservato $\{Y : \mathcal{D} \to \mathcal{E}\}$, assumendo che la distribuzione condizionata del processo Y rispetto al processo X sia tale per cui:

$$\mathfrak{L}_{Y(\mathbf{s})|X} \equiv \mathfrak{L}_{Y(\mathbf{s})|X(\mathbf{s})}, \tag{1.12}$$

ovvero che la legge di $Y(\mathbf{s})$ condizionata rispetto alla realizzazione dell'intero processo X equivalga alla legge di $Y(s)$ condizionata alla sola realizzazione del processo X nel sito $\mathbf{s}$.

Il processo Y potrà avere valori in un sottoinsieme (proprio o improprio) di $\mathbb{R}$ come nel caso presentato nel paragrafo precedente, oppure, più in generale, in un sottoinsieme di $\mathbb{R}^p$ o ancora, come nei casi che tratteremo, $\mathcal{E}$ potrà essere un particolare spazio funzionale infinito dimensionale. Nel seguito utilizzeremo la notazione X_i per indicare $X(\mathbf{s}_i)$.

1.3.2 Markov Random Fields

Vogliamo in un primo tempo fornire un modello per il processo $\{X : \mathcal{D} \to \mathcal{L}\}$, dove $\mathcal{L}$ è un insieme discreto di etichette (*labels*) associate ai siti $\mathbf{s}_i$. Per farlo descriviamo innanzi tutto la struttura che deve avere l'insieme dei siti $\mathcal{S}$, introducendo la nozione di insieme di vicinanza, che ci servirà per descrivere il tipo di dipendenza spaziale vigente tra i siti. Successivamente introdurremo e formalizzeremo l'oggetto di studio, cioè i Markov

Random Fields (MRF), e infine ne specificheremo le proprietà, fornendo una forma per la distribuzione congiunta di $\{X_1, \ldots, X_N\}$.

Struttura di vicinanza

Supponiamo di osservare il processo X in corrispondenza di siti $\mathbf{s}_i$ che rappresentano i punti di una griglia regolare. Per semplicità, inoltre, ci mettiamo nel caso $r = 2$: $\mathcal{S} = \{\mathbf{s}_{(i,j)}, i \in \{1, 2, \ldots, N_1\}, j \in \{1, 2, \ldots, N_2\}\}$. Abbiamo quindi in totale $N := N_1 \times N_2$ realizzazioni del processo X (o di medie del processo su un'area quadrata che ipotizziamo piccola rispetto all'intera area considerata). Le relazioni tra i siti sono espresse da una struttura spaziale detta *insieme di vicinanza*.

Definizione 1.5. *Un insieme di vicinanza $\mathcal{N}_i$ associato al sito i è un insieme di siti che gode delle proprietà seguenti:*

- *un sito non è vicino di sé stesso: $i \notin \mathcal{N}_i$;*
- *la relazione di vicinanza è reciproca: $i \in \mathcal{N}_j \Leftrightarrow j \in \mathcal{N}_i$.*

L'unione di tutti gli insiemi di vicinanza è detta sistema di vicinanza:

$$\mathcal{N} = \{\mathcal{N}_i : i \in \mathcal{S}\}.$$

Per un reticolo piano, un esempio di insieme di vicinanza del sito $\mathbf{s}_i$ è l'insieme di tutti i siti che distano da i meno di un certo raggio R in distanza euclidea. Il più semplice insieme di vicinanza di questo tipo è quello in cui ogni punto interno del reticolo $\mathbf{s}_{(i,j)}$ ha quattro vicini, e cioè i siti adiacenti a nord, sud, est e ovest, di coordinate $\{\mathbf{s}_{(i,j+1)}, \mathbf{s}_{(i,j-1)}, \mathbf{s}_{(i+1,j)}, \mathbf{s}_{(i-1,j)}\}$. Tale struttura è anche detta insieme di vicinanza di primo ordine. La distanza che viene utilizzata per la definizione dell'insieme di vicinanza di primo ordine dipende dalla struttura stessa del reticolo considerato. Nel caso di un reticolo piano, si utilizza la distanza euclidea tra i siti. Nulla vieta, tuttavia, di introdurre insiemi di vicinanza che agiscano attraverso altri tipi di distanza, come per esempio quella geodetica sulla superficie terrestre, se l'insieme $\mathcal{D}$ rappresenta un'area geografica di grandi dimensioni.

Un'altra nozione molto importante per la formalizzazione di un Markov Random Field è il concetto di *clique*.

Definizione 1.6. *Si definisce clique c associata alla coppia $(\mathcal{S}, \mathcal{N})$ ogni insieme di siti che soddisfa una e una sola di queste due proprietà:*

- *c è un insieme che consiste in un solo sito isolato: $c = \{\mathbf{s}_i\}$ per $i \in \{1, \ldots, N\}$;*
- *c è un insieme che consiste in siti tutti mutualmente vicini tra loro, secondo il sistema di vicinanza $\mathcal{N}$: $i, j \in c \Leftrightarrow i \in \mathcal{N}_j$ e $j \in \mathcal{N}_i$.*

Nel caso di un insieme di vicinanza del primo ordine, come quello descritto precedentemente, le sole cliques possibili sono costituite da siti isolati oppure composte di due siti adiacenti, in orizzontale o in verticale. Diamo infine un'ultima definizione che ci servirà in seguito, e cioè la nozione di *potenziale di clique*.

Definizione 1.7. *Un potenziale di ordine k è una funzione di k argomenti simmetrica in tali argomenti.*

Un potenziale di clique $V_c(\mathbf{x})$ è una funzione potenziale i cui argomenti sono i valori di X nei siti appartenenti ad una clique di dimensione k.

La forma del potenziale dipende dal tipo delle variabili considerate, nonché dal tipo di struttura di dipendenza che vogliamo descrivere. Per esempio, se le X_i sono continue un potenziale di ordine 2 può essere semplicemente $x_i x_j$ con $\mathbf{s}_i, \mathbf{s}_j$ appartenenti alla stessa clique.

Markov Random Fields

Definiamo ora esattamente cosa si intende per Random Field:

Definizione 1.8. *Sia $X = \{X_1, \ldots, X_N\}$ una famiglia di variabili aleatorie associate all'insieme dei siti $\mathcal{S}$, tali che ogni variabile aleatoria X_i assuma valori in $\mathcal{L}$. La famiglia X è detta Random Field.*

Ora abbiamo gli strumenti necessari per definire il Markov Random Field:

Definizione 1.9. *Un Random Field X è detto Markov Random Field su $\mathcal{S}$ rispetto al sistema di vicinanza $\mathcal{N}$ se e solo se valgono:*

- $\mathbb{P}(X = \mathbf{x}) > 0 \quad \forall \mathbf{x} = \{x_1, ..., x_N\} \in \mathcal{A} \subset \mathcal{L}^N$ *(positività);*
- $\mathbb{P}(X_i = x_i | \mathbf{x}_{\mathcal{S}\setminus i}) = \mathbb{P}(X_i = x_i | \mathbf{x}_{\mathcal{N}_i})$ *(markovianità).*

Dove $\mathbf{x}_{\mathcal{S}\setminus i}$ indica l'insieme delle etichette dei siti $\mathcal{S}\setminus i$ e $\mathcal{A}$ è un generico sottoinsieme di $\mathcal{L}^N$.

La condizione di positività non costituisce una condizione vincolante. Nel caso in cui tale proprietà non venga rispettata, infatti, basta ridursi a considerare, nell'insieme $\mathcal{A}$, esclusivamente le configurazioni che hanno probabilità non nulla di verificarsi. La markovianità, invece, è la condizione caratterizzante che definisce un MRF. Tale condizione serve a rappresentare il tipo di dipendenza spaziale vigente tra i siti, specificando che le variabili aleatorie $X(\mathbf{s}_i)$ dipendano esplicitamente dall'intera configurazione del campo X solo attraverso i valori assunti nei siti $\mathbf{s}_j$, con $j \in \mathcal{N}_i$

Osserviamo inoltre che tale concetto è una naturale generalizzazione in più dimensioni del concetto di *processo di Markov*.

Gibbs Random Fields

La definizione 1.9 ci suggerisce la possibilità di descrivere un MRF utilizzando la proprietà di markovianità, specificando cioè la forma delle probabilità condizionate $\mathbb{P}(X_i = x_i | \mathbf{x}_{\mathcal{N}_i}) \quad \forall i \in \{1, \ldots, N\}$. Questo è possibile solo se tali probabilità condizionate sono compatibili, cioè portano ad un'unica probabilità congiunta per $\{X_1, \ldots, X_N\}$. Non si possono quindi, in generale, scegliere distribuzioni di probabilità marginali qualsiasi, perché c'è la possibilità che non diano una distribuzione congiunta valida. Per ovviare a questo problema, presentiamo un modo per specificare un MRF direttamente attraverso la probabilità congiunta, come in [Besag, 1974].

Prima di poterlo fare, introduciamo brevemente un altro tipo di RF, e cioè i Gibbs Random Fields. Ricordiamo innanzi tutto il concetto di distribuzione di Gibbs per X, nel caso particolare in cui X sia una famiglia finita di variabili aleatorie a valori in un insieme discreto:

Definizione 1.10. *Una distribuzione di Gibbs per X è una distribuzione di probabilità che assume la forma seguente:*

$$\mathbb{P}(X = \mathbf{x}) = \frac{1}{Z} \exp\{-\frac{1}{T} U(\mathbf{x})\}, \tag{1.13}$$

dove Z è un'opportuna costante di normalizzazione, T è una costante detta temperatura e $U(\mathbf{x})$ è detta funzione energia, e si può esprimere come somma di potenziali di clique, sull'insieme di tutte le possibili cliques $\mathcal{C}$:

$$U(\mathbf{x}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{x}).$$

Dalla forma della distribuzione di probabilità di Gibbs (equazione 1.13), notiamo innanzi tutto che le configurazioni più probabili sono quelle con basse energie. La temperatura T invece controlla la forma della distribuzione: quando T è alta tutte le configurazioni tendono ad essere ugualmente distribuite, mentre per T bassa la distribuzione si concentra attorno alle configurazioni per le quali l'energia globale è minima.

Osserviamo inoltre che, nella definizione della funzione energia $U(\mathbf{x})$ in realtà ogni potenziale di clique tiene conto solo delle variabili che appartengono a tale clique, ovvero il potenziale di clique $V_c(\mathbf{x})$ dipende solo dalla configurazione della clique c e non da tutto $\mathbf{x}$. Possiamo ora dare la definizione di Gibbs Random Field:

Definizione 1.11. *Un Random Field X è detto essere Gibbs Random Field (GRF) su $\mathcal{S}$ rispetto al sistema di vicinanza $\mathcal{N}$ se e solo se X ha distribuzione di Gibbs.*

Il teorema di Hammersley-Clifford

Per ora abbiamo definito due tipi di Random Field: i Markov Random Fields, descritti dalla proprietà locale di markovianità, e Gibbs Random Fields, caratterizzati da una proprietà globale (ovvero la distribuzione di Gibbs). Il teorema che vedremo ora stabilisce un'equivalenza tra le due proprietà. Prima di enunciare il teorema, sono necessarie due assunzioni di base:

- i valori con associata probabilità maggiore di zero per ciascun sito sono in numero finito;
- esiste un valore (il valore $\mathbf{0}$) ammissibile per ogni sito.

La prima proprietà è rilassabile sotto alcune condizioni, tuttavia nel nostro caso risulta sempre soddisfatta in quanto considereremo esclusivamente insiemi di etichette di cardinalità finita. La seconda assunzione invece è una condizione tecnica: garantisce che $\mathbb{P}(\mathbf{0}) > 0$. Sotto tali condizioni, possiamo enunciare il teorema di Hammersley-Clifford:

Teorema 1.1. *X è Markov Random Field su $\mathcal{S}$ rispetto a $\mathcal{N}$ se e solo se X è Gibbs Random Field su $\mathcal{S}$ rispetto a $\mathcal{N}$.*

L'importanza del teorema nelle applicazioni è che permette di specificare un MRF direttamente attraverso la sua distribuzione congiunta, definendo cioè le espressioni dei potenziali di clique $V_c(\mathbf{x})$, in maniera da descrivere il comportamento desiderato.

1.3.3 Campi di Ising

Descriviamo ora un semplice modello di MRF, e cioè il campo di Ising, che verrà utilizzato in seguito per generare dati spazialmente dipendenti, al fine di testare le prestazioni delle tecniche di classificazione spaziale che proporremo nel Capitolo 2. Si tratta di un modello $2D$ a variabili binarie $\mathcal{L} = \{-1, 1\}$ e sistema di vicinanza del primo ordine.

Il modello è stato introdotto per la prima volta in un contesto di modellizzazione di fenomeni magnetici, e si può dunque spiegare in questi termini: consideriamo due

siti $\mathbf{s}_i, \mathbf{s}_j$ appartenenti alla stessa clique (quindi adiacenti). Se $x_i x_j = 1$ significa che i poli sono allineati, dunque si trovano ad un basso livello di energia. Se al contrario $x_i x_j = -1$ i poli sono disallineati, quindi dovranno necessariamente trovarsi ad un alto livello di energia. Dal modello fisico, l'energia totale si scrive:

$$U(\mathbf{x}) = -\frac{J}{2} \sum_{(i,j)\in\mathcal{C}} x_i x_j. \tag{1.14}$$

Dove J è una costante fisica, $\mathbf{x}$ una possibile configurazione e i, j indicano gli indici di due siti appartenenti ad una stessa clique.

Notiamo che l'energia totale è una somma di potenziali di clique di secondo ordine, non sono quindi presenti termini relativi a cliques che contengono siti isolati, in quanto l'energia totale dipenderà esclusivamente dalle disposizioni relative dei singoli siti in gruppi di due.

La probabilità della configurazione $\mathbf{x}$ è, in linea con quanto mostrato precedentemente per i GRF:

$$\mathbb{P}(X = \mathbf{x}) = \frac{1}{Z} \exp\left\{ -\beta \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{N}_i} x_i x_j \right\}, \tag{1.15}$$

dove $\beta = \frac{J}{kT}$ è una costante che caratterizza il campo e riassume tutte le proprietà fisiche del sistema. Si può mostrare inoltre (si veda [Besag, 1974]) che la probabilità condizionata associata al campo si scrive:

$$\mathbb{P}(X_i = x_i | \mathbf{x}_{\mathcal{N}_i}) = \frac{e^{\sum_{j\in\mathcal{N}_i} \beta x_i x_j}}{2\cosh(\sum_{j\in\mathcal{N}_i} \beta x_i x_j)}. \tag{1.16}$$

L'equazione 1.16 può essere utilizzata per campionare dal modello di Ising, mediante metodi MCMC di tipo *Gibbs Sampler*. Si parte da una configurazione casuale, e ad ogni iterazione si trova la realizzazione di X nel sito $\mathbf{s}_i$ campionando da una variabile aleatoria Bernoulliana, con probabilità di successo dipendente solo dai valori assunti dal campo X nei siti vicini, come in 1.16.

Il parametro β regola il grado di dipendenza esistente tra siti vicini. Per valori alti di β (dell'ordine di 1.5, 2) la correlazione spaziale tra siti vicini è più forte, quindi le configurazioni più probabili sono quelle in cui i siti associati alla stessa etichetta tendono ad agglomerarsi; invece per valori bassi di β (dell'ordine di 0.5, 1) la correlazione spaziale è molto più debole, e quindi le configurazioni più probabili sono quelle più sparse. Il comportamento descritto si vede molto bene osservando la Figura 1.4, nella quale mostriamo il campionamento da un campo di Ising ottenuto col metodo MCMC di tipo Gibbs Sampler descritto precedentemente per valori crescenti del parametro β nell'intervallo $[0.25, 1.5]$.

1.3.4 Hidden Random Markov Fields

Abbiamo descritto nei paragrafi precedenti un modo possibile per formalizzare la dipendenza spaziale, attraverso la nozione di Markov Random Field. Nelle applicazioni reali tuttavia, difficilmente le etichette associate ai siti sono osservabili. In generale, anzi, si può solo ipotizzare la presenza di tali etichette, che riassumono determinate caratteristiche di interesse, ma spesso non si riesce a definire con precisione né quante siano le etichette del campo, e nemmeno quali e quante caratteristiche le determinino (si pensi all'esempio dell'area geografica). Per questo motivo, la teoria dei Markov Random

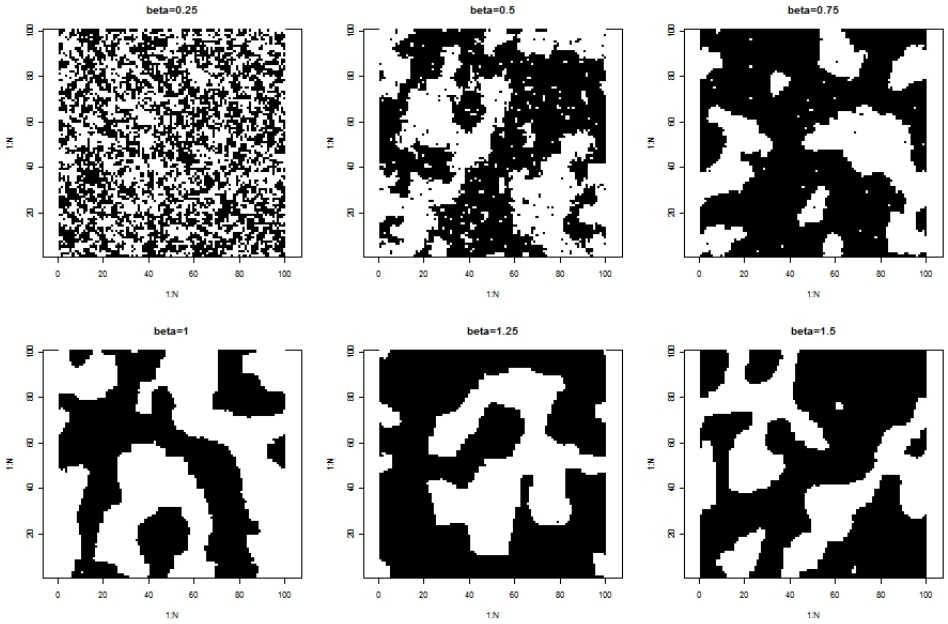


Figura 1.4: Campionamenti dal campo di Ising per diversi valori di β: da sx a dx e dall'alto in basso $\beta = 0.25, 0.5, 0.75, 1, 1.25, 1.5$

Fields viene spesso utilizzata nel contesto più ampio di un modello gerarchico nel quale il campo X è latente, non viene cioè osservato, e quello che si riesce a osservare sono delle altre variabili $Y(\mathbf{s}_i)$ la cui distribuzione dipende esclusivamente dall'etichetta, non osservabile, associata al sito $\mathbf{s}_i$ dal processo X. Nel caso che prenderemo in esame, inoltre, le variabili $Y(\mathbf{s}_i)$ sono funzioni aleatorie, la cui media dipende esclusivamente dall'etichetta latente associata al sito $\mathbf{s}_i$.

Un possibile esempio della situazione che prenderemo in esame è illustrato in Figura 1.5. In Figura, presentiamo una realizzazione di un campo di Ising con $N = 50 \times 50$ siti, e segnale osservato in uno spazio funzionale. Supponiamo, per semplicità, di osservare un segnale con due massimi in corrispondenza dei siti associati all'etichetta $x_i = -1$ (rappresentati in bianco) e un segnale con un massimo in corrispondenza dei siti associati all'etichetta $x_i = +1$ (rappresentati in nero).

Più precisamente, siano $\mathcal{S}$ i punti del reticolo, e $\mathcal{L}$ un insieme discreto di etichette. Sia poi X un Markov Random Field a valori in $\mathcal{L}$, e Y un Random Field a valori in $\mathcal{D}$. Denominiamo con $\mathbf{x} = \{x_1, \ldots, x_N\} \in \mathcal{L}^N$ una possibile configurazione di etichette, e con $\mathbf{y} = \{y_1, \ldots, y_N\} \in \mathcal{D}^N$ una possibile configurazione osservata negli N siti di $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\}$.

Supponiamo infine che esista una legge di probabilità (denominata distribuzione di emissione) che associa l'osservazione y_i nel sito $\mathbf{s}_i$ all'attribuzione del sito ad una certa etichetta $x_i = l$ del campo latente, $l \in \mathcal{L}$, cioè che esista la legge:

$$\mathfrak{L}_{Y(\mathbf{s}_i)|X=\mathbf{x_i}} \equiv \mathfrak{L}_{Y(\mathbf{s}_i)|X(\mathbf{s}_i)=x_i}. \tag{1.17}$$

L'ipotesi principale che descrive la struttura di un HRMF è che la distribuzione condizionata del processo $Y(\mathbf{s}_i)$ osservabile nel sito $\mathbf{s}_i$ condizionatamente alla realizzazione del campo X, dipenda esclusivamente dalla realizzazione del campo latente in quel sito, ovvero dall'etichetta l associata al sito $\mathbf{s}_i$, indipendentemente dalla realizzazione dl campo nei siti vicini. La struttura di dipendenza spaziale, quindi, è presente esclusivamente

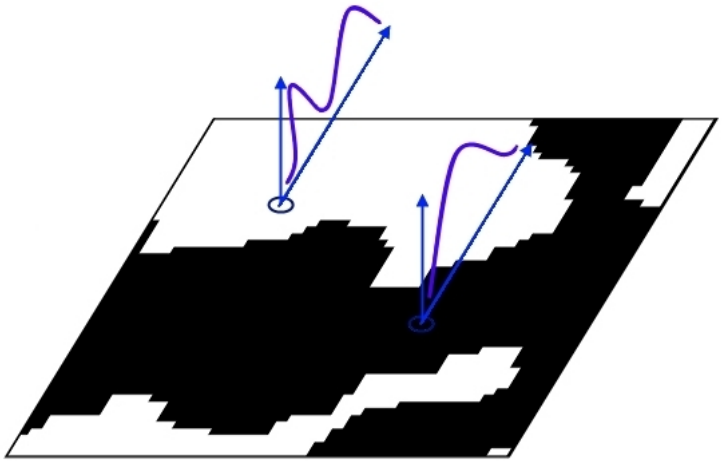


Figura 1.5: Illustrazione del concetto di HRMF nel caso di campo latente di Ising e segnale osservato in uno spazio funzionale

a livello del campo Markoviano latente X, e condizionatamente alla realizzazione di tale campo, le variabili $Y(\mathbf{s}_i)$ sono tra loro indipendenti.

In generale, anche se il campo X ha valori in un insieme discreto, la distribuzione di emissione che lega il campo osservato Y al campo latente di etichette X può essere continua e, per esempio, multivariata. I parametri della distribuzione di emissione, nonché lo spazio sul quale è definita, dipendono dalla situazione in esame. Poiché siamo intenzionati a studiare casi in cui il dato y_i appartiene ad uno spazio funzionale, non specifichiamo una forma parametrica per tale distribuzione, ma ipotizziamo soltanto che:

$$Y_i(t)|(X_i = l) = F_l(t) + \epsilon_i(t), \tag{1.18}$$

nella quale $\epsilon_i(t)$, sono funzioni aleatorie normali indipendenti e identicamente distribuite, dette funzioni di errore:

$$\epsilon_i(t) \overset{\text{iid}}{\sim} N(0, \Sigma),$$

dove con 0 indichiamo la funzione identicamente nulla e con $\Sigma = \Sigma(t, s)$ una funzione di autocovarianza. La 1.18 implica che il dato funzionale y_i associato al sito $\mathbf{s}_i$ sia la somma di due termini: una funzione aleatoria $F_l(t)$ la cui distribuzione dipende esclusivamente dal valore dell'etichetta x_i associata al sito, ed è quindi comune a tutti i dati associati alla stessa etichetta del campo latente, ed una funzione di errore $\epsilon_i(t)$.

Nel corso dei prossimi capitoli ci proponiamo quindi di sviluppare delle tecniche di classificazione dei dati y_i che permettano di ricostruire la disposizione delle etichette del campo latente. Vogliamo quindi risalire alla media delle funzioni $F_l(t)$, al fine di fornire, per ognuna delle classi di etichette, un unico dato rappresentante della classe, eliminando le funzioni di errore $\epsilon_i(t)$.

Capitolo 2

Procedura di classificazione per dati funzionali spazialmente correlati

2.1 Introduzione al problema di classificazione

Nel corso di questo capitolo, proporremo ed analizzeremo delle tecniche che permettono l'esplorazione e classificazione di dati funzionali spazialmente dipendenti.

Supponiamo quindi di avere a disposizione un set di dati funzionali, ovvero di poter osservare, per ogni unità statistica associata al sito $\mathbf{s}_i \in \mathcal{S}$, l'evoluzione di un determinato fenomeno lungo una variabile continua, per esempio il tempo, nell'intervallo (t_{min}, t_{max}). Supponiamo inoltre che il fenomeno osservato sia una quantità scalare, indicata con la variabile aleatoria Y. Ovviamente, non siamo in grado di osservare l'andamento continuo del fenomeno, ma esclusivamente il valore che assume la variabile Y in corrispondenza di alcuni istanti di tempo $t_1 < t_2 < ... < t_J$. I dati che ci proponiamo di analizzare saranno quindi costituiti da osservazioni successive del fenomeno di interesse.

Per analizzare tali dati, è possibile trattare le osservazioni $\{y(t_1), ..., y(t_J)\}$ come realizzazione di un vettore aleatorio di dimensione J, ovvero $\mathbf{Y} = \{Y(t_1), ..., Y(t_J)\}$, ed utilizzare le tecniche classiche di analisi multivariata.

Un'altra possibilità è considerare le osservazioni $Y(t_j)$ come realizzazioni di un processo stocastico $\mathcal{Y} = \{Y(t), t \in [t_{min}, t_{max}]\}$, che chiameremo funzione aleatoria. L'analisi statistica da delle osservazioni $\{y(t_1), ..., y(t_J)\}$ richiede dunque due passi successivi: in un primo momento è necessario sviluppare tecniche (dette di *smoothing*) per passare dalle osservazioni discrete $\{y(t_j)\}_i \quad j = 1, ..., J$ (dove i indica l'unità statistica) alla funzione $y_i(t)$. Una volta ottenute le funzioni $y_i(t)$, una per ogni unità statistica, si passa a effettuare le analisi desiderate tenendo conto del fatto che, ora, i dati non sono vettori ma realizzazioni della funzione aleatoria $\mathcal{Y}$.

Il secondo approccio si sta sempre più diffondendo in letteratura, e presenta molti vantaggi nella rappresentazione dei dati quando si suppone che i dati *whithin–unit*, cioè le osservazioni puntuali $\{y(t_j)\}_i \quad j = 1, ..., J$ riferite alla stessa unità statistica, presentino un rapporto segnale rumore molto alto (si veda [Ke and Wang, 2001], [Sangalli et al., 2009]). Uno dei vantaggi di considerare il dato come una funzione è la possibilità di esplorare molte più caratteristiche legate all'evoluzione del fenomeno, come per esempio calcolarne le derivate. Nella nostra analisi sceglieremo quindi questo secondo approccio, immaginando di osservare realizzazioni discrete di un fenomeno continuo, e di dover dunque ricostruire ed analizzare dati di tipo funzionale.

Nel seguito, per semplicità di notazione, indicheremo la funzione aleatoria in esame con Y, e le realizzazioni di tale funzione con y_i, senza specificare che si tratta di funzioni.

Nel nostro caso poi, ipotizziamo che gli indici delle unità statistiche siano dei punti (o siti) in uno spazio metrico $\mathcal{D}$, in corrispondenza dei quali osserviamo il dato funzionale. Per quanto riguarda la struttura dell'insieme $\mathcal{S}$ dei siti in corrispondenza del quale osserviamo i dati, ipotizziamo che il tipo di struttura spaziale a disposizione sia una griglia regolare $\mathcal{S} = \{(i,j), i,j \in \{1,2,\cdots,N\}\}$. Supponiamo quindi che il set di dati così costituito non sia indipendente, ma i dati presentino una dipendenza spaziale; in particolare ipotizziamo che la dipendenza spaziale tra dati funzionali agisca solo attraverso un campo latente di etichette associate ai siti del reticolo, che supponiamo essere un MRF. In altri termini, ipotizziamo di osservare la realizzazione di un HRMF con distribuzioni di emissione infinito dimensionale e campo latente discreto.

L'obiettivo dell'analisi è classificare i dati funzionali, sfruttando le informazioni provenienti dalla dipendenza spaziale tramite una tecnica di classificazione opportunamente sviluppata. Si tratta di una tecnica esplorativa, come la maggior parte delle tecniche sviluppate nell'ambito della Functional Data Analysis (FDA). Fare inferenza statistica su dati di tipo funzionale, infatti, presenta difficoltà legate alla dimensione elevata dello spazio in cui sono definiti i dati: un dato funzionale, in generale, è rappresentato da una funzione in uno spazio infinito dimensionale, della quale possiamo avere esclusivamente un numero finito di osservazioni, rendendo l'inferenza statistica estremamente difficoltosa.

Più precisamente, il problema consiste nell'associare ad ogni sito $s \in \mathcal{S}$ un'etichetta, o label $l \in \mathcal{L}$, in modo da suddividere regioni di siti i cui dati siano tra loro omogenei ed associarle alla stessa classe. Vogliamo risolvere un problema di classificazione non supervisionata, nel quale il numero di etichette, e la loro natura (ovvero l'eventuale significato fisico associato ai gruppi), sono incogniti, e dunque le etichette stesse sono associate in modo arbitrario (ovvero interessa il risultato della suddivisione in k classi, non il particolare valore numerico assegnato all'etichetta).

In un problema come quello brevemente descritto, di classificazione non supervisionata, l'insieme delle etichette $\mathcal{L}$ è un insieme discreto di cardinalità k, dove k è il numero di cluster nei quali si vogliono dividere i dati di partenza. Per ora supporremo che k sia noto; in seguito affronteremo problemi più generali, nei quali nemmeno il numero di cluster è noto, e svilupperemo quindi una tecnica per scegliere quanti cluster utilizzare.

Infine, per descrivere il problema in un contesto spaziale, è necessario scegliere come misurare la distanza tra i siti $d(\cdot,\cdot)$, in modo che essa sia definita nello spazio cui appartiene la regione $\mathcal{S}$. In generale, può essere utilizzata una qualsiasi metrica, che deve essere scelta a seconda del problema da analizzare. La distanza d può descrivere l'effettiva distanza geografica tra i siti, o alternativamente può tenere conto di un'eventuale anisotropia del campo latente, qualora questa informazione fosse disponibile.

A titolo di esempio, decliniamo i concetti enunciati in precedenza in un caso semplice, ovvero la ricostruzione di un'immagine a due colori (per esempio bianco e nero) di dimensioni $N_1 \times N_2$ (in pixel), partendo da osservazioni funzionali che in ogni punto dell'immagine dipendono da tali colori: in tal caso, l'insieme delle etichette è $\mathcal{L} = \{0,1\}$, l'insieme dei siti è costituito da un reticolo $\mathcal{S} = \{(i,j), i,j \in \{1,2,\cdots,N\}\}$, e la distanza d è la distanza euclidea in $\mathbb{R}^2$.

2.2 Tecniche per trattare la dipendenza spaziale

L'obiettivo che ci prefiggiamo è sfruttare opportunamente la dipendenza spaziale esistente tra i dati col fine di ricostruire il campo latente delle etichette.

In generale, immaginiamo che se i dati associati ai k gruppi dal processo Y provengono da distribuzioni stocasticamente lontane tra loro, un possibile approccio sia quello di classificare i dati così come sono per mezzo di algoritmi noti, senza tenere conto della loro collocazione nello spazio (cioè non utilizzando nell'analisi le informazioni riguardanti i siti $\mathbf{s}$). Questa idea è sconsigliabile per tre principali ragioni: innanzi tutto, se si ipotizza che i dati siano dipendenti, è scorretto utilizzare algoritmi che li classifichino sfruttando l'ipotesi di indipendenza, poiché tale ipotesi è violata.

In secondo luogo, nel caso di dati spazialmente dipendenti, la posizione nel reticolo di tali dati costituisce un'informazione in più che può essere sfruttata per la classificazione, rendendo il risultato più accurato poiché utilizza tutte le informazioni in nostro possesso. Infatti, qualora la struttura di dipendenza spaziale fosse legata ad un MRF, per trovare l'etichetta x_i associata al sito $\mathbf{s}_i$ sarebbe consigliabile utilizzare non solo l'informazione su x_i proveniente dall'osservazione y_i nel sito in questione, ma anche i valori del processo nei siti vicini, poiché anch'essi contengono informazioni riguardanti l'etichetta x_i.

Infine, in molte applicazioni reali, ci si deve confrontare con problemi di grandi dimensioni (immaginiamo che l'insieme dei siti sia un reticolo fitto su una vasta area geografica), per i quali il numero totale di siti N è molto grande. La classificazione dell'insieme di tutti gli N dati funzionali costituirebbe quindi un problema di natura computazionale.

Queste prime considerazioni suggeriscono l'idea sulla quale si basa l'algoritmo che utilizzeremo per la classificazione. Gli algoritmi di clustering per dati funzionali più diffusi sono basati sull'ipotesi che i dati siano tra loro approssimativamente indipendenti. Volendo utilizzare tali algoritmi per i nostri dati, che invece risultano dipendenti tra loro a causa del campo spaziale latente, ci proponiamo di creare un datasetdi sintesi di dimensioni inferiori costituito da dati che supponiamo essere indipendenti, a partire dai dati originali e sfruttando le informazioni spaziali. L'ottenimento del dataset di dimensioni inferiori, costituito da dati rappresentativi della dipendenza spaziale latente si basa sulla seguente:

Definizione 2.1. *Dato uno spazio metrico $(\mathcal{S}, d(\cdot,\cdot))$, e un insieme finito di punti di $\mathcal{S}$, $\mathcal{C} = \{c_1, \cdots, c_n\} \subset \mathcal{S}$ o centri, si dice diagramma (o tassellazione) di Voronoi associata a $\mathcal{C}$ la partizione di $\mathcal{S}$ che associa ad ogni $c_j \in \mathcal{C}$ una regione $V(c_j|\mathcal{C})$ in modo tale che tutti i punti di $V(c_j|\mathcal{C})$ siano più vicini a c_j che ad ogni altro punto in $\mathcal{C}$, ovvero:*

$$V(c_j|\mathcal{C}) = \{x \in \mathcal{S} : d(x, c_j) \leq d(x, c_i) \quad \forall c_i \in \mathcal{C}, \quad i \neq j\}.$$

A questo punto, possiamo descrivere l'idea sulla quale si basa lo sfruttamento della dipendenza spaziale nell'algoritmo di classificazione: per ridurre il dataset in modo da ottenere dati tra loro indipendenti effettuiamo una tassellazione di Voronoi dell'insieme $\mathcal{S}$ a partire da un insieme di centri $\{c_1, ..., c_n\}$ generati casualmente in maniera uniforme su $\mathcal{S}$. Ad esempio, se scegliamo $\mathcal{S}$ come un reticolo piano quadrato di dimensioni $N1 \times N_2 = 400 \times 400$ (quindi $N = 400^2$), e $n = 100$, un possibile risultato è quello mostrato in Figura 2.1.

Ad ogni tassello così ottenuto associamo un dato funzionale, calcolando la media di tutti i dati associati ai siti del tassello con un opportuno kernel, in modo da pesare di più i dati associati a siti più vicini al centro e di meno i dati associati a quelli più lontani.

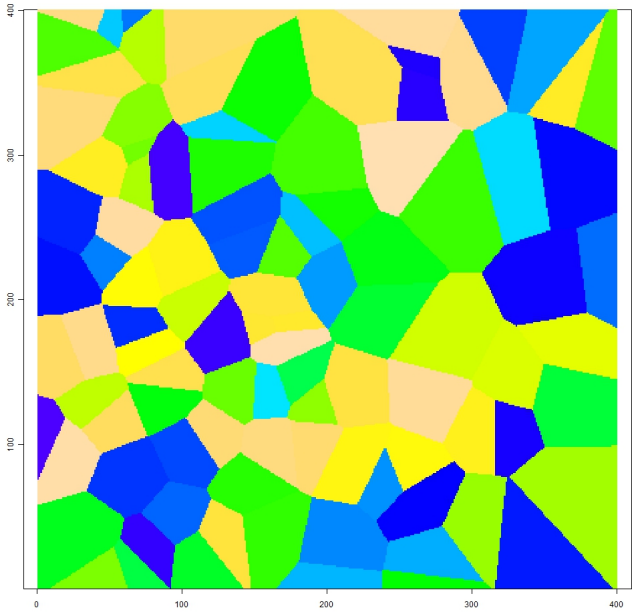


Figura 2.1: Una possibile tassellazione di Voronoi a partire da centri generati casualmente in maniera uniforme su un'area quadrata costituita da 400×400 siti

In particolare, calcoliamo un dato di sintesi $\tilde{y}_k(t)$ per ogni tassello k come media pesata dei dati associati al tassello, con opportuni pesi w_i:

$$\tilde{y}_k(t) = \frac{\sum_{\mathbf{s}_i \in V(c_k|\mathcal{C})} w_i y_i(t)}{\sum_{\mathbf{s}_i \in V(c_k|\mathcal{C})} w_i}. \tag{2.1}$$

Se la dimensione del tassello rispecchia il grado di dipendenza spaziale tra i siti, è possibile ipotizzare che dati appartenenti allo stesso tassello siano tra loro simili, ovvero che l'etichetta associata dal campo latente sia la stessa , e dunque che la loro media (a netto di un kernel dipendente dalla posizione) sia un buon rappresentante della media della funzione aleatoria $F_l(t)$. Inoltre in questo caso, l'informazione spaziale associata al dato funzionale è già contenuta nel corrispondente dato di sintesi. Per tali ragioni si può supporre che gli n dati rappresentativi così ottenuti $\{\tilde{y}_1(t), ..., \tilde{y}_n(t)\}$ presentino una dipendenza molto più debole rispetto ai dati originari, preservandone le caratteristiche importanti. Il nuovo dataset di sintesi rispetto al quale effettuare clustering funzionale è quindi composto semplicemente dai dati medi su ogni tassello.

2.3 Tecniche di classificazione

L'algoritmo procede classificando i dati di sintesi in k cluster, con l'utilizzo delle tecniche di classificazione non supervisionata note dalla letteratura sui dati funzionali. Le tecniche di clustering utilizzate possono essere diverse a seconda del tipo di dati in esame, per adattarsi in modo flessibile al problema considerato.

Un metodo possibile per classificare i dati è effettuare una riduzione dimensionale proiettando i dati rispetto ad un'opportuna base finito dimensionale (per esempio la base di Fourier), per poi classificare con tecniche multivariate i vettori di coefficienti associati alle funzioni di base tramite k–medie. Se la base utilizzata è ortogonale, infatti,

la distanza $L^2([t_{min}, t_{max}])$ tra le proiezioni di due funzioni sul sottospazio generato dalla base è proporzionale alla distanza euclidea tra i vettori dei coefficienti corrispondenti.

Un secondo metodo proposto per la classificazione è una riduzione dimensionale tramite componenti principali funzionali (FPCA), per poi effettuare la classificazione non supervisionata per mezzo di k–medie sugli scores. L'idea è esattamente la stessa del caso precedente, cioè quella di proiettare i dati su una base dello spazio $L^2([t_{min}, t_{max}])$ per poi classificare i vettori composti dai coefficienti della proiezione. Quello che cambia rispetto all'approccio precedente è che, ora, la base utilizzata non è stabilita a priori, ma costruita a partire dai dati stessi.

Ricordiamo brevemente che effettuare un'analisi di tipo FPCA su di un dataset consiste nel cercare una base dello spazio funzionale $L^2([t_{min}, t_{max}])$ composta da funzioni che rappresentino le direzioni lungo le quali la variabilità dei dati è maggiore (per un'analisi più approfondita si veda [Ramsay and Silverman, 2005]).

Più precisamente, sia Y una funzione aleatoria, $Y : \Omega \to L^2([t_{min}, t_{max}])$. Definiamo innanzi tutto la media $\mu(t)$ e la funzione di autocovarianza $\Sigma(t, s)$ della funzione Y, analogamente al caso vettoriale:

$$\mu(t) = \mathbb{E}[Y(t)] \quad \forall t \in [t_{min}, t_{max}], \tag{2.2}$$

$$\Sigma(t, s) = \mathrm{Cov}(Y(t), Y(s)) = \mathbb{E}[(Y(t) - \mu(t))(Y(s) - \mu(s))] \quad \forall (t, s) \in [t_{min}, t_{max}]^2. \tag{2.3}$$

La prima funzione della base (prima componente principale) è quindi quella funzione $\Phi_1(t)$ che individua la direzione in $L^2([t_{min}, t_{max}])$ lungo la quale la varianza dei dati è massima:

$$\Phi_1(t) = \underset{\Phi \in L^2([t_{min}, t_{max}]) : \|\Phi\| = 1}{\operatorname{argmax}} \mathrm{Var}(\langle \Phi, Y \rangle), \tag{2.4}$$

dove $\langle \cdot, \cdot \rangle$ indica il prodotto scalare in $L^2([t_{min}, t_{max}])$, ovvero:

$$\langle \Phi, \Psi \rangle = \int_{t_{min}}^{t_{max}} \Phi(t)\Psi(t)\, dt. \tag{2.5}$$

Ricordando che il valore del prodotto scalare tra due elementi di uno spazio vettoriale è la lunghezza della proiezione ortogonale di un elemento sull'altro, possiamo dare un'interpretazione più precisa della prima componente principale. Nella 2.4, infatti, massimizzare l'espressione $\mathrm{Var}(\langle \Phi, Y \rangle)$ equivale a trovare quella particolare funzione $\Phi_1(t)$ per la quale la varianza delle proiezioni dei dati y_i su Φ_1 è massima. La funzione $\Phi_1(t)$, quindi, permette di descrivere i dati nel miglior modo possibile in uno spazio ad una sola dimensione poiché una volta nota $\Phi_1(t)$, per descrivere $y_i(t)$ basta indicare il valore di $\langle \Phi, y_i \rangle$, che viene denominato *score* di y_i associato alla prima componente principale. In sintesi, la prima componente principale non è altro che una funzione che identifica il sottospazio monodimensionale nel quale la variabilità dei dati è massima.

Analogamente, la seconda componente principale è quella funzione $\Phi_2(t)$ che massimizza la varianza della proiezione di Y su tutte le funzioni del sottospazio ortogonale a Φ_1:

$$\Phi_2(t) = \underset{\Phi \in L^2([t_{min}, t_{max}]) : \|\Phi\| = 1, \langle \Phi, \Phi_1 \rangle = 0}{\operatorname{argmax}} \mathrm{Var}(\langle \Phi, Y \rangle), \tag{2.6}$$

la terza componente principale è la funzione che massimizza la variabilità nel sottospazio ortogonale a Φ_1 e Φ_2 e così via.

Si può mostrare che le componenti principali così ottenute non sono altro che le autofunzioni dell'operatore di covarianza $V_\Sigma : L^2([t_{min}, t_{max}]) \to L^2([t_{min}, t_{max}])$ così definito:

$$(V_\Sigma \Phi)(t) = \int_{t_{min}}^{t_{max}} \Sigma(t, s)\Phi(s)\, ds, \tag{2.7}$$

ovvero che risolvano l'equazione:

$$\lambda_i \Phi_i(t) = \int_{t_{min}}^{t_{max}} \Sigma(t,s)\Phi_i(s)\, ds. \tag{2.8}$$

Inoltre, gli autovalori associati alle autofunzioni $\Phi_1(t), \Phi_2(t), ...$ sono tali che $\lambda_1 \geq \lambda_2 \geq ...$, in modo tale che la prima componente principale Φ_1 sia l'autofunzione associata all'autovalore di modulo massimo λ_1, e così via.

Una volta trovata la base ortonormale delle componenti principali (o autofunzioni), si proiettano i dati sulla base così definita (arrestando l'espansione all'ordine q), e si procede con la classificazione multivariata dei vettori degli scores dei dati, cioè i coefficienti delle proiezioni dei dati funzionali lungo le prime q autofunzioni. Si tratta di un metodo di classificazione generale, basato su un approccio di tipo data-driven, poiché la base scelta per proiettare i dati è indotta dai dati stessi. Tale metodo può quindi essere applicato a dati di tipo molto diverso. Gli scores oggetto della classificazione non sono altro che i coefficienti associati alle funzioni della base composta dalle componenti principali funzionali.

In questo caso si può mostrare (si veda [Ferraty and Vieu, 2006]) che fare k–medie sugli scores è equivalente a fare k–medie funzionale in base ad una semimetrica indotta dalle componenti principali. Ricordiamo che una *semimorma* su uno spazio F è una funzione $\|\cdot\| : F \to [0, +\infty)$ che soddisfa le seguenti proprietà:

- $\forall(\lambda, x) \in \mathbb{R} \times F, \quad \|\lambda x\| = |\lambda| \|x\|$;
- $\forall(x, y) \in F \times F, \quad \|x + y\| \leq \|x\| + \|y\|$.

Si tratta quindi di una norma, fatta eccezione per la proprietà di annullamento. Non si richiede cioè $\|x\| = 0 \Rightarrow x = 0$.

Analogamente, una *semimetrica* d può essere definita come una metrica per la quale tuttavia $d(x, y) = 0 \not\Rightarrow x = y$. Nel caso delle componenti principali funzionali, se scegliamo di considerare le prime q componenti, la semimetrica che andiamo a considerare è costituita dalla distanza L^2 tra i dati proiettati sullo spazio q-dimensionale costituito dalle prime q autofunzioni:

$$d_q^{PCA}(y_1, y_2) = \sqrt{\sum_{l=1}^{q} \left(\int_{t_{min}}^{t_{max}} [y_1(t) - y_2(t)]\Phi_l(t)\, dt \right)^2}. \tag{2.9}$$

La semimetrica 2.9 è esattamente la distanza euclidea tra i vettori q-dimensionali degli scores delle due funzioni y_1 e y_2, che sono definiti come le proiezioni delle funzioni lungo le direzioni identificate dalle prime q autofunzioni, cioè le quantità:

$$\int_{t_{min}}^{t_{max}} y(t)\Phi_l(t)\, dt \quad l = 1, ..., q \tag{2.10}$$

Un'osservazione in più va fatta per quanto riguarda la scelta di quali e quante componenti principali utilizzare. Siamo in effetti interessati non tanto alle direzioni lungo le quali la variabilità dei dati è maggiore, quanto alle direzioni lungo le quali gli scores associati ai dati si separano meglio nei gruppi. Nel caso in cui k sia piccolo e si osservi in maniera evidente la presenza di più mode nelle distribuzioni degli scores, scegliamo di prendere in considerazione quelle componenti che permettano una migliore classificazione nei k gruppi, scegliendo direttamente tra gli istogrammi degli scores quelli che presentano una più marcata plurimodalità.

Questa osservazione si può utilizzare soprattutto per valori di k piccoli, tipicamente per $k = 2$ o 3, poiché se facciamo crescere il numero di cluster è sempre più difficile osservare una k–modalità evidente nelle distribuzioni univariate di ciascuno score. In tal aso, si può procedere ricercando la k–modalità nelle distribuzioni degli scores in più dimensioni. Nel caso in cui la plurimodalità non sia comunque visibile o evidente, bisognerà ragionare esclusivamente sulla percentuale di variabilità spiegata, nonché sull'interpretabilità delle autofunzioni delle componenti principali, per scegliere quelle che permettano una migliore rappresentazione dei dati.

Una volta classificati in uno dei modi descritti i dati di sintesi, ad ogni sito appartenente ad un medesimo tassello viene associata l'etichetta del dato di sintesi corrispondente. Se il tassello contiene al suo interno siti associati alla medesima etichetta del campo latente l'effetto della procedura di tassellazione e del calcolo del dato medio è esclusivamente la riduzione della variabilità dei dati di partenza, senza perdita di alcuna informazione sull'etichetta associata ai punti del tassello. Se invece il tassello contiene siti associati ad etichette diverse del campo latente, mediando i dati associati ai siti appartenenti al tassello mischieremo informazioni provenienti da distribuzioni distinte, rendendo più difficile la classificazione del dato medio risultante; in questo caso compiremmo comunque un errore classificando insieme punti con etichette di partenza diverse.

Poiché la divisione in tasselli è aleatoria, e non è detto che segua perfettamente i confini tra le zone associate ad etichette diverse del campo latente,. Di conseguenza, questa procedura verrà ripetuta più volte, cambiando ogni volta la tassellazione di partenza e dunque anche il set di dati di sintesi da classificare.

Alla fine dell'algoritmo scegliamo di associare ogni sito del reticolo ad un cluster in base ai risultati delle classificazioni di ogni run dell'algoritmo, per mezzo di una procedura di voto di maggioranza: se nella maggior parte dei casi un sito è stato associato ad una data etichetta, il sito sarà classificato con quella etichetta.

2.4 Scelta del numero di cluster

La procedura che abbiamo proposto nei paragrafi precedenti, sfrutta la dipendenza spaziale per ridurre il dataset in analisi e procede con la classificazione dei dati di sintesi. Le tecniche proposte per la classificazione, tuttavia, si basano tutte sull'ipotesi di essere a conoscenza della cardinalità dell'insieme $\mathcal{L}$ delle etichette, che viene utilizzata nell'algoritmo per fissare il numero di cluster k.

Dalle considerazioni fatte all'introduzione del capitolo, tuttavia, possiamo osservare che, in generale, il numero di etichette diverse non è noto a priori, quindi non può essere utilizzato come un dato dell'algoritmo. A questo punto, quindi, è necessario sviluppare delle tecniche che ci permettano di scegliere tra diversi valori di k, quello che porta ad un migliore risultato della classificazione finale.

Per risolvere il problema posto, è necessario innanzi tutto definire cosa si intende per buona classificazione. Ricordiamo infatti che abbiamo posto il problema come classificazione non supervisionata, nella quale cioè le etichette di partenza non sono note, e non possiamo quindi calcolare indici di misclassificazione. Dobbiamo quindi cercare una misura della bontà del risultato ottenuto che non dipenda dalle etichette del campo latente, essendo queste non note.

Ripercorrendo lo schema che abbiamo proposto per risolvere il problema di classificazione, possiamo fare la seguente considerazione: abbiamo detto che la procedura di tassellazione e conseguente calcolo dei dati di sintesi è aleatoria, e viene ripetuta più

volte per ottenere la stima finale. Nulla tuttavia ci assicura che due diverse classificazioni effettuate a partire da diverse tassellazioni di Voronoi diano dei risultati tra loro coerenti per quanto riguarda la divisione nei diversi cluster. Se la classificazione effettuata è fittizia, o deriva esclusivamente dalla particolare tassellazione ottenuta, ci aspettiamo, anzi, che i risultati ottenuti alle diverse iterazioni dell'algoritmo siano tra loro molto diversi. Se invece la classificazione ottenuta rispecchia la reale suddivisione dei dati in gruppi all'interno dei quali i dati hanno caratteristiche comuni, ci aspettiamo che le diverse classificazioni ottenute partendo da diverse tassellazioni diano risultati tra loro comparabili, e che tali risultati differiscano esclusivamente nelle regioni di confine tra i diversi gruppi, a causa della diversa disposizione dei tasselli nello spazio $\mathcal{D}$.

Una misura della diversità tra una classificazione e l'altra può essere data dalle frequenze puntuali di assegnazione ai diversi cluster, così calcolate: fissato un sito $\mathbf{s}_i$ e un cluster h ($i \in \{1, 2, ..., N\}, h \in \{1, 2, ..., k\}$), chiamiamo frequenza di assegnazione del sito $\mathbf{s}_i$ al cluster h la quantità

$$f_{ih} = \frac{M_{ih}}{M},$$

dove ricordiamo che M è il numero totale di iterazioni fatte, mentre indichiamo con M_{ih} il numero di volte in cui il sito $\mathbf{s}_i$ è stato classificato nel cluster h.

Riformulando quanto abbiamo osservato precedentemente in termini di frequenze puntuali di assegnazione, quindi, possiamo supporre che la classificazione effettuata sia buona e corrisponda alla reale struttura di raggruppamento dei dati se le frequenze di assegnazione dei siti sono vicine ai due valori 0 o 1 (un sito viene associato sempre allo stesso cluster); si deve invece supporre che la classificazione finale sia fittizia o non corrisponda alla situazione reale se osserviamo che le frequenze di assegnazione di moli siti hanno valori intermedi (un sito viene associato ogni volta ad un cluster diverso).

A questo proposito, una quantità che possiamo prendere in considerazione come indicatore globale della qualità di una classificazione è l'entropia media (introdotta come indice di bontà nell'ambito della ricostruzione di immagini in [Leung and Lam, 2002]) definita come segue:

Definizione 2.2. *Sia k il numero di cluster fissato per la classificazione, effettuata col metodo precedentemente descritto. Siano $f_{i1}, f_{i2}, ..., f_{ik}$ le frequenze di assegnazione ai k cluster relative a un sito $\mathbf{s}_i$. Definiamo entropia della classificazione del sito $\mathbf{s}_i$ la quantità:*

$$\eta(k)_i = -\sum_{h=1}^{k} f_{ih} \log f_{ih}. \tag{2.11}$$

Definiamo entropia media della classificazione la quantità

$$\eta(k) = \frac{\sum_{\mathbf{s}i \in \mathcal{S}} \eta(k)_i}{N}. \tag{2.12}$$

Osserviamo innanzi tutto che si tratta di quantità sempre positive, che variano tra 0 e $\log k$, dove k indica il numero di cluster. Osserviamo poi che l'entropia calcolata in un sito $\mathbf{s}_i$ è una quantità che assume valore massimo quando $f_{ih} = \frac{1}{k} \quad \forall h$, cioè quando il sito viene classificato uno stesso numero di volte in tutti i possibili cluster, ed è uguale a zero per $f_{ih^*} = 1, f_{ih} = 0 \quad \forall h \neq h^*$, cioè quando il sito viene classificato ogni volta nello stesso cluster h^*. Un'entropia media vicina a 0, quindi, può essere utilizzata come indice globale di bontà di una classificazione; quindi l'entropia media è l'indice che proponiamo di utilizzare per la scelta del numero di cluster k.

Ricordando che l'entropia media assume valori in intervalli diversi al variare del numero di cluster, per confrontare valori ottenuti in classificazioni con diversi valori di k è dunque necessario normalizzare il valore ottenuto, dividendolo per $\log k$ in modo da confrontare tra loro quantità che assumono valori nello stesso intervallo $[0, 1]$. Una volta effettuata la normalizzazione, scegliamo k^* che minimizzi il valore dell'entropia media normalizzata:

$$k^* = \underset{k}{\operatorname{argmin}} \frac{\eta(k)}{\log k}. \tag{2.13}$$

Per trovare tale valore k^*, effettuiamo la classificazione per diversi valori di k, e per ognuna delle classificazioni ottenute calcoliamo il valore dell'entropia media, per poi scegliere il minimo.

A questo proposito, c'è da fare un'ultima precisazione: se la classificazione non è fittizia, ci aspettiamo che l'entropia sia alta nelle regioni di confine tra un cluster e l'altro. Se aumentiamo il numero di cluster, tuttavia, tali regioni aumenteranno di conseguenza, quindi c'è da aspettarsi un leggero aumento dell'entropia. Considerando anche questo fenomeno, possiamo pensare di scegliere il numero di cluster in corrispondenza del quale osserviamo un gomito (cioè un aumento molto ripido) nel grafico dell'entropia in funzione del numero di cluster.

2.5 Schema dell'algoritmo proposto

In sintesi, il metodo proposto per la classificazione è un algoritmo basato sul campionamento ripetuto, che consiste nel ripetere M volte i seguenti passi:

1. Tassellazione;
2. Calcolo dei dati medi;
3. Classificazione dei dati medi.

Alla fine delle M run dell'algoritmo,si calcola la stima finale del campo latente delle etichette associando ogni sito al cluster al quale quel sito è stato associato con frequenza maggiore nel corso delle M run.

I parametri da fissare per il funzionamento dell'algoritmo, escludendo il numero di cluster k, che può essere scelto con il metodo della minimizzazione dell'entropia proposto nella Sezione 2.4, sono quindi il numero di run M ed il numero di tasselli del diagramma di Voronoi n. Il problema della scelta del numero di tasselli del diagramma di Voronoi n verrà discusso in maniera dettagliata nel Capitolo 3, nel quale effettueremo la classificazione di dataset sintetici con l'algoritmo proposto, e analizzeremo le prestazioni dell'algoritmo, in termini di rate di misclassificazione, al variare del numero di tasselli.

Per quanto riguarda, invece, la scelta del numero di run M, osserviamo che, in linea teorica, più run vengono effettuate, più la stima finale ottenuta è da considerarsi attendibile, poiché ottenuta a partire da un numero elevato di classificazioni indipendenti. D'altra parte, aumentando eccessivamente il parametro M, si va incontro a problemi legati all'elevato ordine computazionale dell'algoritmo. La scelta del parametro dipende quindi dal problema preso in esame, e dai mezzi a disposizione per l'analisi. Nel seguito del lavoro, sceglieremo di effettuare 50 o 100 run dell'algoritmo, poiché nel nostro caso tale scelta sembra sufficiente per ottenere delle stime che non risentano in maniera eccessiva dei risultati ottenuti nelle singole run.

In Figura 2.2 mostriamo il diagramma di flusso della procedura proposta.

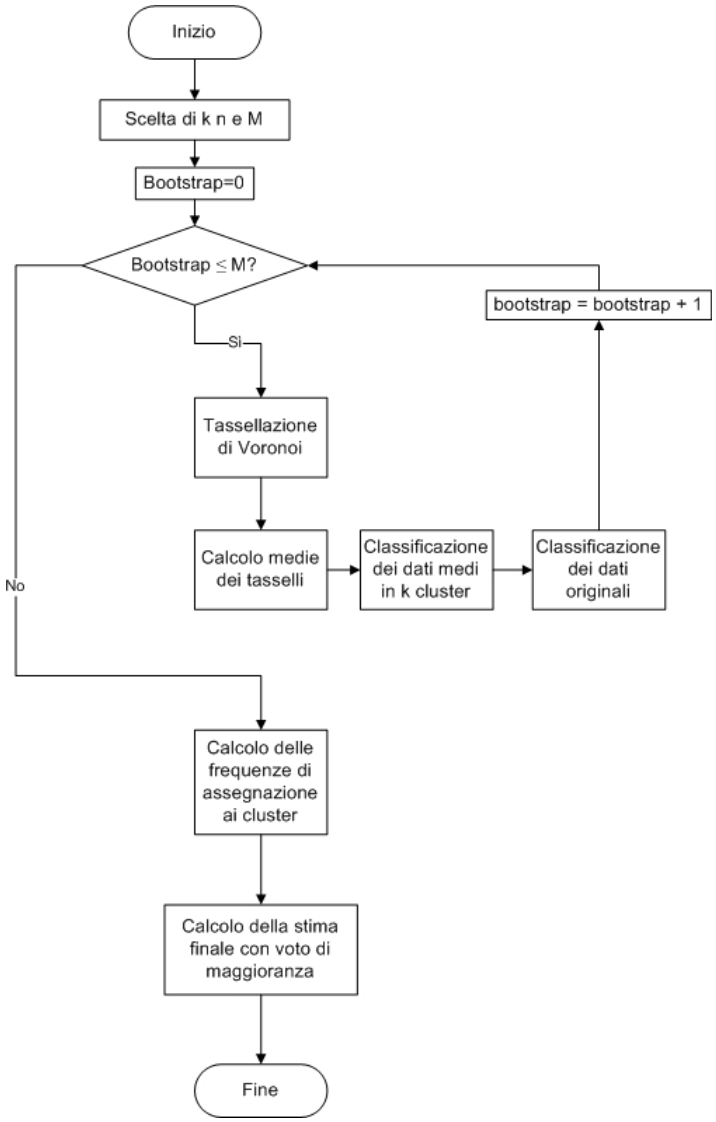


Figura 2.2: Diagramma di flusso dell'algoritmo di classificazione proposto

Capitolo 3

Studi di simulazione

3.1 Simulazioni da un modello Hidden Random Markov Field

Nel corso di questo capitolo, analizzeremo le tecniche per l'esplorazione e classificazione di dati funzionali descritte nel Capitolo 2. Lo scopo di tale analisi è uno studio delle prestazioni delle tecniche introdotte su dati simulati (dei quali quindi si conosce la distribuzione spaziale delle etichette) al fine della loro validazione e utilizzo su casi reali. Nel corso di questo capitolo, quindi, innanzi tutto proporremo dei modi per generare dati funzionali con dipendenza spaziale che possano avere le stesse caratteristiche dei dati reali ai quali si vorranno applicare le tecniche proposte. Una volta generati i dati, secondo i modelli di dipendenza spaziale visti nel Capitolo 1, utilizzeremo l'algoritmo descritto nel Capitolo 2 per effettuare la classificazione, adattandone i vari passaggi alla situazione in esame. Mostreremo quindi i risultati della classificazione di tali dati in termini di errore di misclassificazione delle etichette note, per verificare le prestazioni delle tecniche introdotte, e per confrontarle con alcune tecniche di classificazione per dati indipendenti note.

3.1.1 Descrizione del problema

Al fine di effettuare gli studi di simulazione presentati nella prima parte del capitolo, generiamo un set di dati sintetici a partire da un modello di dipendenza spaziale descritto nel Capitolo 1, ovvero un Hidden Markov Random Field: supponiamo di osservare dati funzionali emessi da un Random Markov Field latente. Supponiamo inoltre che il campo Markoviano latente sia un campo di Ising, cioè un MRF piano e isomorfo su un reticolo di dimensioni $N_1 \times N_2$ pixel, le cui etichette sono variabili binarie (ovvero l'insieme $\{-1, 1\}$, che verrà associato nelle visualizzazioni ai due colori $\{bianco, nero\}$).

Nella seconda parte del capitolo complicheremo tale modello di generazione dei dati introducendo un secondo grado di dipendenza spaziale tra i dati funzionali, che non risulteranno più essere indipendenti condizionatamente alle etichette.

Ricordiamo che siamo interessati esclusivamente ad una classificazione dei dati in insiemi o cluster distinti, in modo che i dati di ogni insieme siano quelli generati da etichette distinte. Per ora, quindi, non ci poniamo il problema di fare inferenza sui parametri del campo latente, né sul tipo di distribuzione di emissione.

Dal punto di vista metodologico, è necessario ora declinare l'algoritmo generale presentato nel Capitolo 2 al caso in esame, tramite alcune scelte specifiche. Prima di tutto, supponiamo che il numero di cluster sia noto, e sia $k = 2$. In secondo luogo, le

unità statistiche sono dati funzionali associati ad un reticolo di punti su un piano. La distanza che utilizziamo per la tassellazione di Voronoi è quindi quella euclidea su $\mathbb{R}^2$.

Resta da definire il metodo da utilizzare per la classificazione dei dati. In questo primo caso,opereremo una riduzione dimensionale del dato utilizzando le componenti principali funzionali (FPCA), per poi effettuare la classificazione non supervisionata per mezzo di k–medie sugli scores. Per quanto riguarda la scelta delle componenti da utilizzare, scegliamo di considerare le componenti che permettono una migliore classificazione nei due gruppi, selezionando direttamente tra gli istogrammi degli scores quelli che presentano una più marcata bimodalità, come descritto nella Sezione 2.3.

Nei paragrafi seguenti riportiamo una serie di simulazioni effettuate per testare in questo contesto la bontà dei risultati forniti dall'algoritmo descritto nel Capitolo 2, sotto diverse ipotesi di modello.

3.1.2 Modello di generazione dei dati

Vogliamo analizzare il funzionamento dell'algoritmo descritto per mezzo di diverse simulazioni. Per prima cosa abbiamo bisogno di simulare dei dati funzionali che abbiano il tipo di dipendenza spaziale descritto nel Paragrafo 3.1.1.

Volendo considerare il caso a due cluster, generiamo innanzi tutto un campo di Ising di dimensione $N_1 \times N_2$, in cui a ciascuno dei siti è associata un'etichetta scelta nell'insieme $\{-1, 1\}$; scegliamo poi il parametro di riferimento $\beta = 2$, in modo da ottenere una situazione nella quale i due cluster siano tra loro ben distinti.

Una volta ottenuta la realizzazione del campo di Ising, cioè la disposizione nello spazio delle etichette, dobbiamo simulare il dato funzionale. In particolare, scegliamo di simulare funzioni periodiche in uno spazio finito dimensionale, che è quello generato dalle prime p funzioni della base di Fourier sull'intervallo $[0, 1]$. Richiediamo la periodicità delle funzioni per una questione di coerenza con i dati climatici che analizzeremo nel Capitolo 4. La scelta di tale spazio funzionale permette di simulare i dati in modo molto semplice: basta infatti scegliere i p coefficienti associati alle funzioni della base e creare, a partire dai coefficienti, il dato funzionale.

Dovremo quindi generare p coefficienti per ogni punto del campo di Ising di partenza (o sito), ovvero $N_1 \times N_2$ vettori $\mathbf{c}_i$ di dimensione p indipendenti condizionatamente alle etichette (dove i indica il sito del campo di Ising X). La legge che genera i coefficienti $\mathbf{c}_i$ in ogni sito dovrà quindi dipendere esclusivamente dall'etichetta x_i associata a quel sito. Ipotizziamo inoltre che solo la media $\boldsymbol{\mu}_{x_i}$ del vettore di coefficienti dipenda dall'etichetta del sito, e che la matrice di covarianza sia costante. Supponiamo infine che tali vettori siano generati da una normale multivariata e che i coefficienti associati ad un singolo punto siano tra loro indipendenti (matrice di covarianza diagonale e costante per ogni sito $\Sigma_i = \Sigma = \sigma^2 I$):

$$\mathbf{c_i}|x_i \sim N_p(\boldsymbol{\mu}_{x_i}, \sigma^2 I).$$

3.1.3 Piano di simulazione e scelta dei parametri

Dal momento che le etichette possibili per ogni sito sono solo due, per simulare i dati abbiamo bisogno di definire i seguenti parametri:

- dimensioni del campo delle etichette $N_1 \times N_2$;
- numero di coefficienti della base di Fourier p;
- varianza dei coefficienti σ^2;

- vettori delle medie dei coefficienti dei due gruppi $\boldsymbol{\mu_{-1}}$ e $\boldsymbol{\mu_{+1}}$.

Occorre poi scegliere i metodi di classificazione dei dati, il numero di iterazioni dell'algoritmo ed infine il numero di tasselli del diagramma di Voronoi da utilizzare per ridurre il dataset.

Nelle simulazioni effettuate scegliamo di generare un campo di Ising su di un reticolo quadrato, $N_1 = N_2 = \sqrt{N} = 100$, $p = 5$, e $\sigma^2 = 1$ oppure 0.5. Per quanto riguarda la distribuzione spaziale delle etichette, la stessa realizzazione del campo di Ising verrà utilizzata in tutte le simulazioni presentate nella prima parte di questo capitolo, ed è quella mostrata in Figura 3.1.

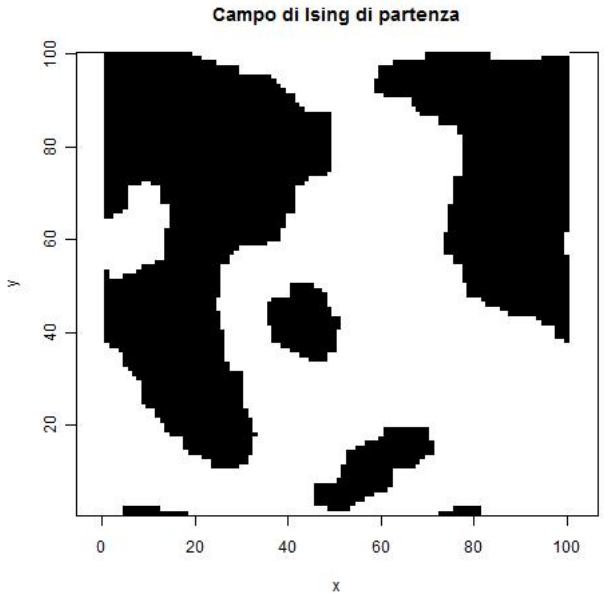


Figura 3.1: Realizzazione di un campo di Ising con $\beta = 2$

Per quanto riguarda invece le medie dei vettori dei coefficienti dei due gruppi, scegliamo di farle variare considerando otto diverse situazioni, da una prima situazione nella quale le medie dei due gruppi sono praticamente coincidenti, all'ultima nella quale si osservano distintamente i due diversi cluster. I vettori delle medie dei coefficienti dei diversi casi sono riassunti in Tabella 3.1, mentre i dati funzionali associati ai vettori medi scelti (rispettivamente per $\sigma^2 = 1$ e $\sigma^2 = 0.5$) sono riassunti nelle Figure 3.2 e 3.3. I grafici mostrano i dati funzionali associati ai valori medi dei coefficienti (linea

	$(\mu_{-1})_1$	$(\mu_{-1})_2$	$(\mu_{-1})_3$	$(\mu_{-1})_4$	$(\mu_{-1})_5$	$(\mu_1)_1$	$(\mu_1)_2$	$(\mu_1)_3$	$(\mu_1)_4$	$(\mu_1)_5$
Caso 0	1	1.75	1.75	0	0	1	1.75	1.75	0	0.25
Caso 1	1	1.75	1.75	0	0	1	1.75	1.75	0.25	0.25
Caso 2	1	1.75	1.75	0	0	1	1.75	1.75	0.5	0.5
Caso 3	1	1.75	2	0	0	1	1.75	1.50	0.5	0.5
Caso 4	1	2	2	0	0	1	1.5	1.50	0.5	0.5
Caso 5	1	2	2.25	0	0	1	1.5	1.25	0.75	0.75
Caso 6	1	2	2.5	0	0	1	1.5	1	1	1
Caso 7	1	2	2.75	0	0	1	1.5	0.75	1.25	1.25

Tabella 3.1: Medie delle distribuzioni dei coefficienti associate alle due etichette

continua), e quelli risultanti prendendo i valori medi puntuali più o meno due volte la deviazione standard puntuale (linea tratteggiata), per ciascuno dei due gruppi.

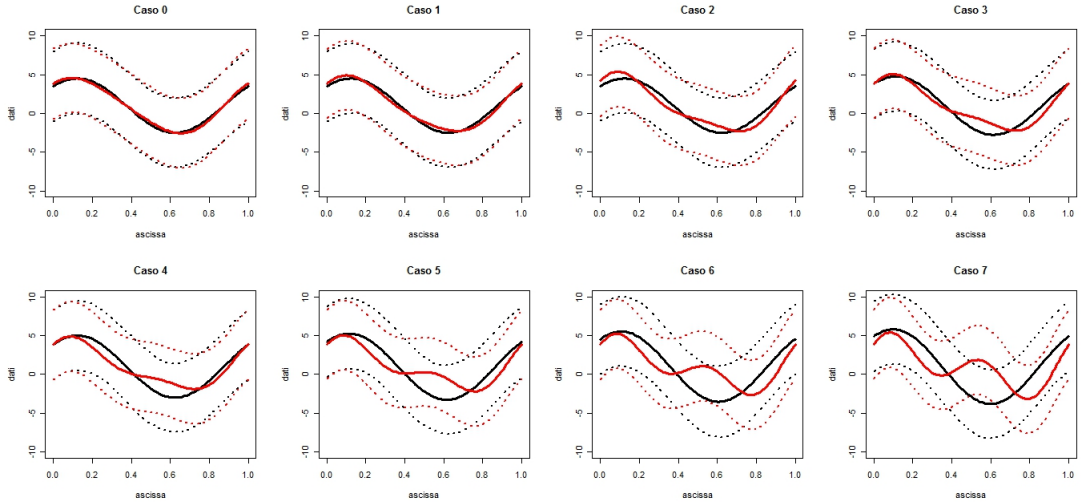


Figura 3.2: Dati funzionali ottenuti con i coefficienti dati dai valori medi $(-)$ e dai valori medi puntuali $\pm$ due volte la deviazione standard puntuale $(--)$ per $\sigma^2 = 1$

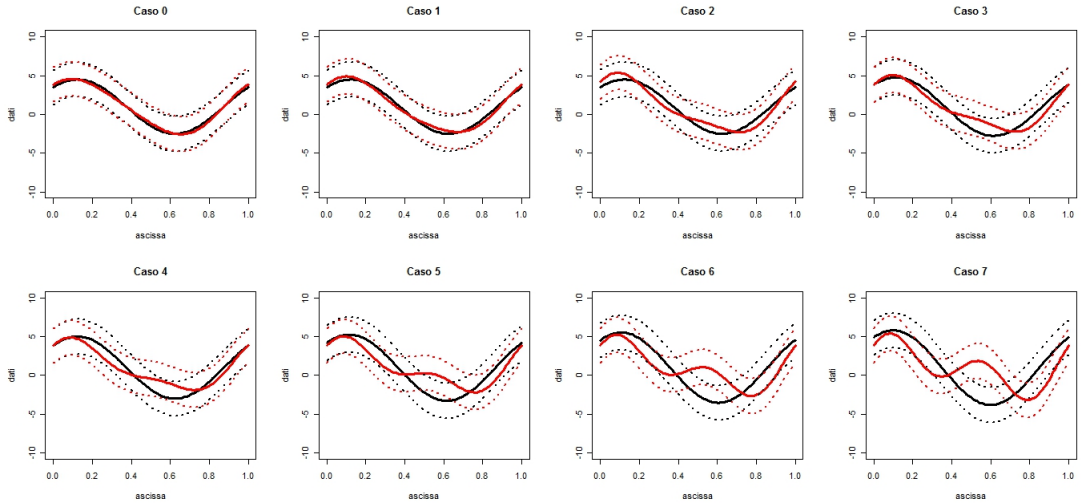


Figura 3.3: Dati funzionali ottenuti con i coefficienti dati dai valori medi $(-)$ e dai valori medi puntuali $\pm$ due volte la deviazione standard puntuale per $\sigma^2 = 0.5$

Per la classificazione dei dati così simulati sono stati scelti tre differenti metodi:

1. *simple clustering*: classificazione dei dati originali tramite riduzione dimensionale (FPCA) e k–medie sugli scores (corrisponde a classificare i dati non considerando le informazioni provenienti dalla posizione nello spazio);

2. *spatial clustering*: classificazione dei dati originali tramite l'algoritmo descritto nel Capitolo 2, ovvero calcolo dei dati di sintesi su una tassellazione di Voronoi e loro clustering e k–medie funzionale con semimetrica indotta dalle FPC;

3. *trivial clustering*: classificazione banale di tutti i dati in un solo cluster.

I metodi simple clustering e trivial clustering costituiscono i due casi limite dello spatial clustering per particolari valori del numero di tasselli n: il simple clustering corrisponde al caso in cui si scelga un numero di tasselli equivalente al numero di siti, $n = N$; il trivial clustering, al contrario, corrisponde al caso in cui si scelga di effettuare la classificazione di un unico tassello che ricopre l'intero reticolo, $n = 1$. Questi due metodi servono da confronto per testare il comportamento dell'algoritmo presentato. Ci aspettiamo infatti che l'algoritmo proposto, che sfrutta la dipendenza spaziale dei dati, funzioni meglio della classificazione dei dati originali, che non tiene conto delle informazioni spaziali, e meglio di un classificatore banale che associa tutti i dati ad un unico cluster. Infatti, come già osservato, se i dati presentano una dipendenza spaziale, per associare al sito $\mathbf{s}_i$ un'etichetta x_i, è possibile sfruttare le informazioni su x_i che provengono dai dati osservati in un intorno del sito $\mathbf{s}_i$, come viene fatto nel metodo dello spatial clustering.

Se, al contrario, i dati sono tra loro indipendenti, i valori del processo Y in un intorno del sito $\mathbf{s}_i$ sono indipendenti dl valore di Y nel sito, non possono quindi contribuire in nessun modo alla stima dell'etichetta x_i, e l'unico metodo utilizzabile per classificare i dati è quello del simple clustering.

Scegliamo poi di lanciare l'algoritmo di classificazione con $M = 100$ iterazioni. L'ultimo parametro che resta da fissare a questo punto è il numero n dei tasselli del diagramma di Voronoi. Come spiegato nella descrizione dell'algoritmo, la dimensione del tassello dovrebbe rispecchiare il grado di dipendenza spaziale presente nell'immagine di partenza. Se i tasselli sono troppo grandi, a fronte di una cospicua riduzione della variabilità dei dati, molti tasselli conterranno siti associati ad etichette diverse e il risultato sarà distorto; al limite, il valore massimo che può assumere la superficie di un tassello è la taglia dell'intero reticolo, e ritroviamo il trivial clustering (massima distorsione, minima varianza). Se invece i tasselli sono troppo piccoli, rischiamo di ridurre troppo poco la variabilità dei dati, non sfruttando adeguatamente le informazioni spaziali; al limite, il tassello più piccolo possibile coincide con un unico sito del reticolo, e ritroviamo quindi la classificazione dei dati originali (minima distorsione, massima varianza). Tale comportamento prende il nome in letteratura di *bias–variance trade–off*. Aumentando il numero di tasselli, il dataset composto dai dati di sintesi si avvicina al dataset originale. La distorsione è bassa ma la variabilità del dataset da classificare è molto alta. Se, al contrario, diminuiamo il numero di tasselli, man mano aumenterà la distorsione del dataset analizzato, a fronte di una cospicua riduzione della variabilità del dataset da classificare. Supponiamo, quindi, che ci sia un numero di tasselli *ottimo*, per il quale si introduce una distorsione del dataset, ma la varianza è stata ridotta, migliorando complessivamente i risultati della classificazione.

Per poter studiare il trade–off distorsione-varianza, scegliamo quindi di far variare il numero dei tasselli, in modo da cambiarne la dimensione. Il parametro che prenderemo in considerazione è una sorta di densità di tasselli rispetto alle dimensioni dell'immagine

$\rho = \frac{n}{N^2}$. La classificazione banale e quella dei dati originali rappresentano rispettivamente i valori $\rho = \frac{1}{N^2} \simeq 0$ e $\rho = 1$. Per la nostra analisi, scegliamo inoltre di far variare il parametro ρ tra i valori $0.005, 0.01, 0.015, 0.02, 0.025$.

3.1.4 Risultati delle simulazioni

Analizziamo ora i risultati delle analisi di classificazione su dataset che abbiamo generato. Innanzi tutto, presentiamo per qualche caso particolare i risultati ottenuti, osservando i dati di partenza, i dati medi, gli scores associati alle prime componenti principali, la scelta delle componenti da utilizzare per la classificazione e i risultati della classificazione stessa.

Prendiamo in esame due diversi tipi di dati iniziali, ovvero dati simulati nel caso 1 e nel caso 5 riassunti in Figura 3.2, entrambi con parametro $\sigma^2 = 1$. Nel primo caso, i dati dei due gruppi sono tra loro molto simili, mentre nel secondo caso sono visivamente diversi. Osservando il grafico dei dati medi dei tasselli (per il valore $\rho = 0.005$), in Figura 3.4, tale differenza si vede molto bene. Inoltre soprattutto dal grafico di destra (quello del caso in cui le differenze nelle medie dei due coefficienti sono più marcate) vediamo come in effetti i dati di sintesi siano raggruppati intorno alle medie delle due distribuzioni, e come siano presenti dati di sintesi dall'andamento funzionale intermedio ai due gruppi, conseguenza del fatto che alcuni tasselli ricoprono zone miste. Il fatto che tali aspetti si possano vedere distintamente osservando i dati è conseguenza del fatto che calcolando le medie dei dati sui tasselli, la variabilità è diminuita. In Figura 3.5, infatti, mostriamo 50 dati estratti casualmente dal dataset originario[1]. Notiamo dalla Figura che non è possibile distinguere i due gruppi, né nel caso 1 né nel caso 5, poiché la variabilità è troppo alta.

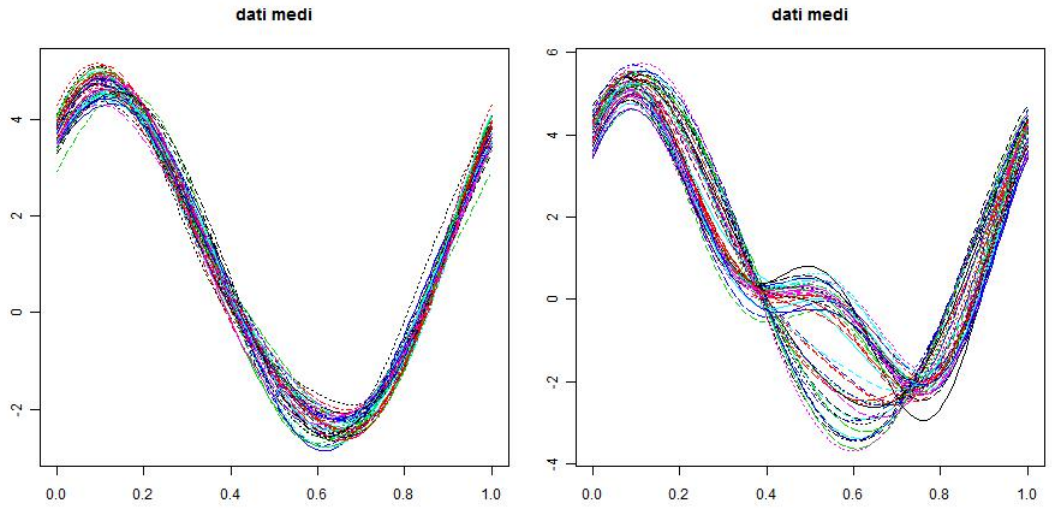


Figura 3.4: Dati di sintesi del caso 1 (sinistra) e 5 (destra) per $\sigma^2 = 1$, $\rho = 0.005$

Una volta ottenuti i dati medi, poi, l'algoritmo procede effettuando una riduzione dimensionale con componenti principali funzionali sui dati medi. Gli istogrammi degli scores delle prime 4 componenti principali sono presentati in Figura 3.6. Nel caso in cui le medie sono più vicine, non sembra che nessuno degli istogrammi mostrati

[1] Date le dimensioni elevate del dataset, scegliamo di rappresentare i dati di origine esclusivamente estraendone un campione casuale della stessa numerosità del campione dei dati di sintesi

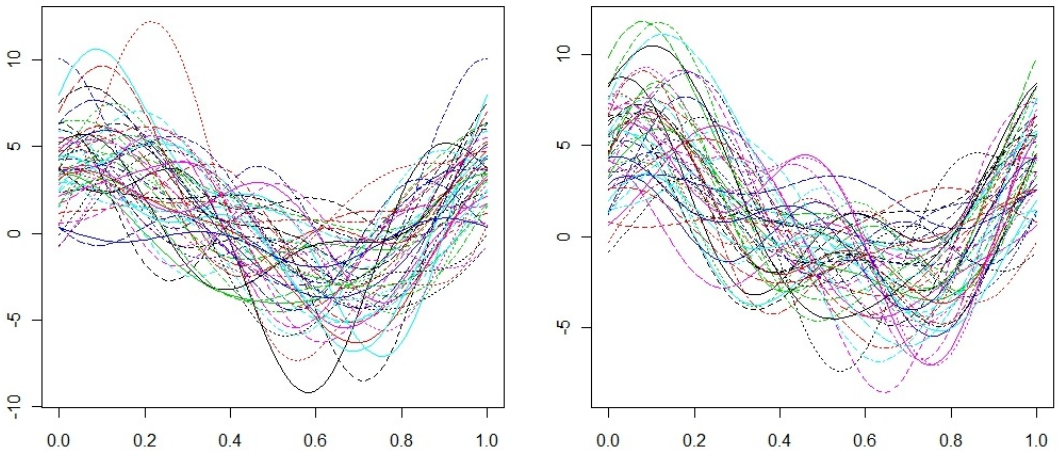


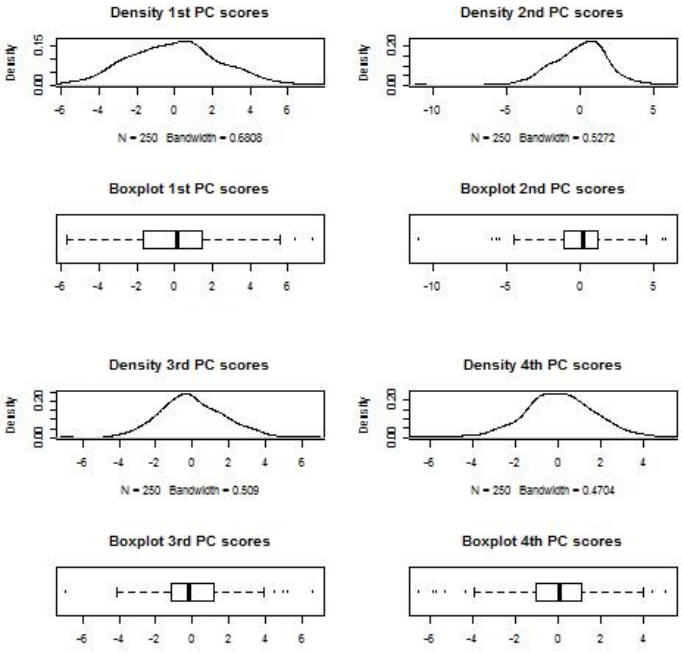
Figura 3.5: Estrazione casuale di 50 dati originali del caso 1 (sinistra) e 5 (destra) per $\sigma^2 = 1$

presenti bimodalità, proprio a causa del fatto che non ci sono differenze marcate tra i due gruppi. Nel secondo caso preso in esame, invece, c'è un'evidente bimodalità nella prima componente principale.

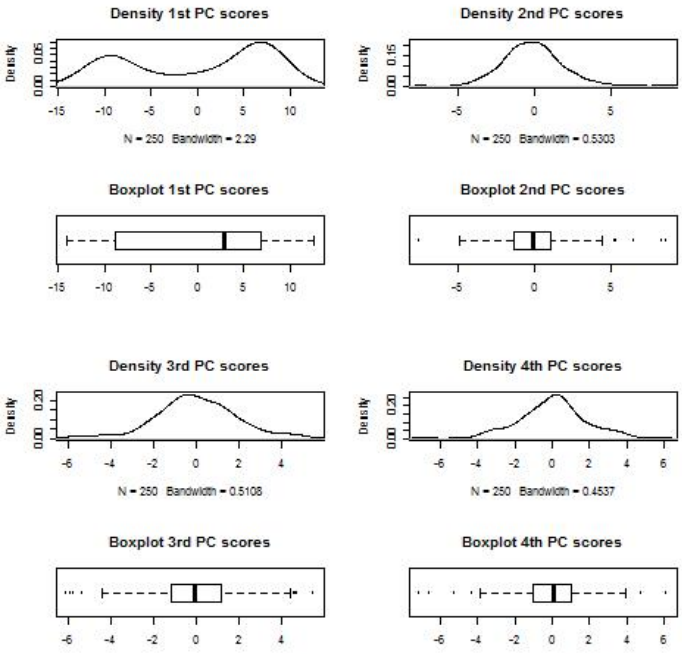
Prima di procedere nella scelta delle componenti principali da utilizzare per la classificazione, quindi, analizziamo le componenti principali ottenute, cercando di visualizzare quale parte di variabilità sono in grado di spiegare. Nelle Figure 3.7 e 3.8 osserviamo le curve medie dei dati (la linea continua) alla quale abbiamo sommato (linea formata da $+$) o sottratto (linea formata da $-$) le autofunzioni moltiplicate per l'autovalore loro corrispondente, mentre le autofunzioni stesse sono mostrate nelle Figure 3.9 e 3.10. I grafici dei dati medi ai quali vengono sommate o sottratte le autofunzioni sono, in generale, un modo per visualizzare l'effetto di una particolare componente: da un questi grafici si vede infatti come viene perturbata la media dei dati dalle diverse componenti.

Fissiamoci su una componente principale, per esempio la prima. I dati i cui scores lungo la prima componente principale sono positivi, si disporranno lungo la linea formata da segni $+$, mentre quelli associati a scores negativi si disporranno lungo la linea formata da segni $-$. Se ci sono zone in cui l'effetto di perturbazione dell'autofunzione sulla media è più marcato (le due linee positiva e negativa sono più distanti) significa che quella componente sta evidenziando la diversità dei dati in quelle zone. Questo aspetto, nei dati che stiamo analizzando, si nota esclusivamente osservando il grafico relativo alla prima componente principale del caso 5, cioè il primo pannello in Figura 3.8, dove osserviamo che la componente identifica esattamente la zona nella quale i dati associati ai due gruppi sono tra loro più diversi.

Per quanto riguarda i grafici relativi a tutte le altre componenti, invece, le tre linee sono molto vicine in tutte le zone del grafico. Questo sta a significare che la variabilità dei dati (descritta dagli autovalori relativi alle diverse componenti ottenute da FPCA sui dati di sintesi) è molto bassa, cosa che del resto si vede osservando i dati di sintesi stessi, e dal grafico non si riesce a distinguere il comportamento dei dati nei diversi gruppi. Per questo motivo mostriamo anche l'andamento delle sole autofunzioni, nelle Figure 3.9 e 3.10. Per quanto riguarda il caso 1, vediamo che la prima autofunzione cambia segno nella zona del grafico nella quale i dati dei due gruppi si distanziano, anche se, a causa della bassa variabilità, la cosa non si vede nel grafico superiore. Per questo motivo



(a) Caso 1



(b) Caso 5

Figura 3.6: Densità e boxplot degli Scores delle prime 4 componenti principali per i casi 1 e 5

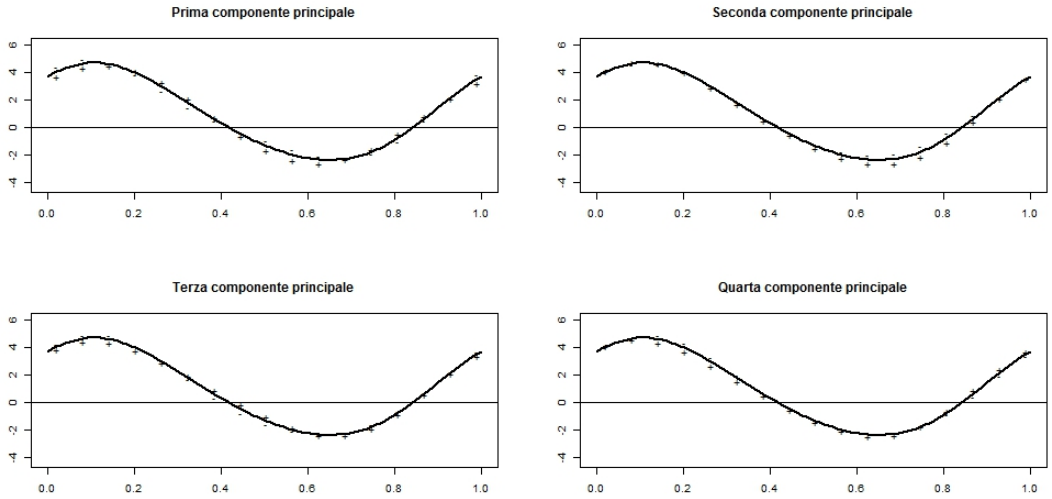


Figura 3.7: Curve medie e perturbazione della media aggiungendo (+) o sottraendo (-) le prime quattro autofunzioni moltiplicate per il relativo autovalore nel primo caso, $\sigma^2 = 1$

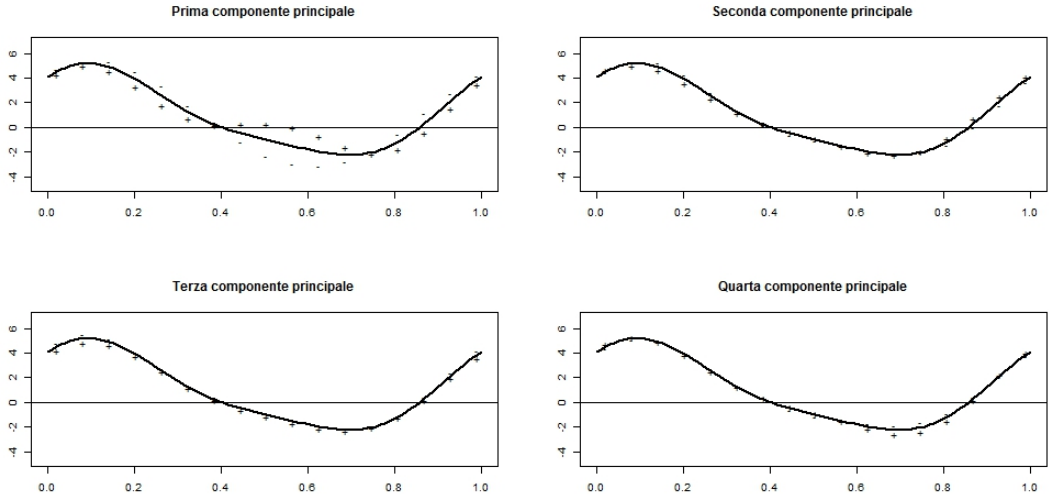


Figura 3.8: Curve medie e perturbazione della media aggiungendo (+) o sottraendo (-) le prime quattro autofunzioni moltiplicate per il relativo autovalore nel quinto caso, $\sigma^2 = 1$

scegliamo di mantenere la prima componente principale in entrambi i casi. Per una rappresentazione più completa del dato, poi, manteniamo anche la seconda componente principale e procediamo con la classificazione.

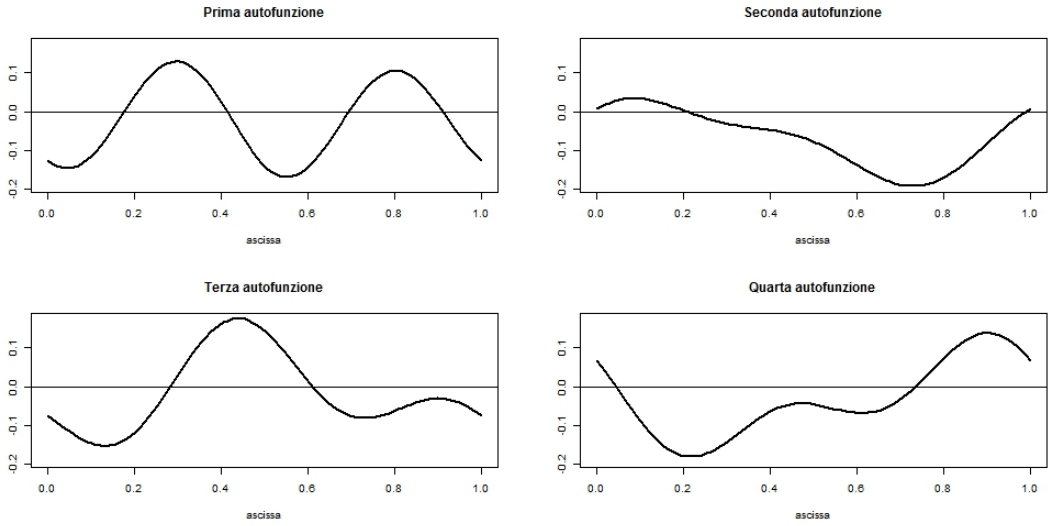


Figura 3.9: Prime quattro autofunzioni nel primo caso, $\sigma^2 = 1$

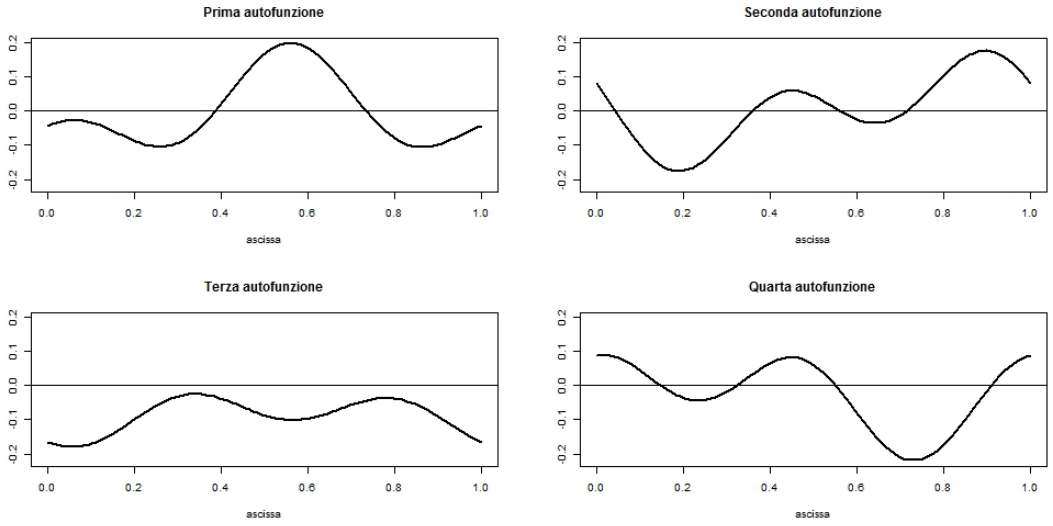
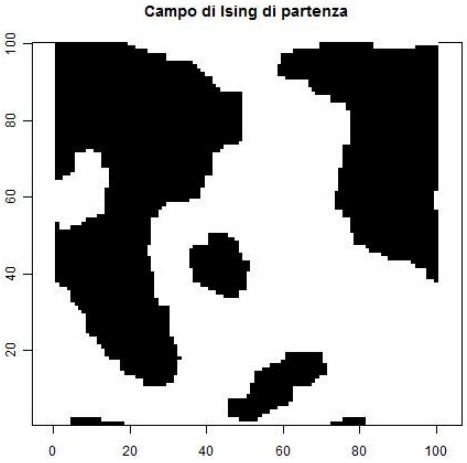


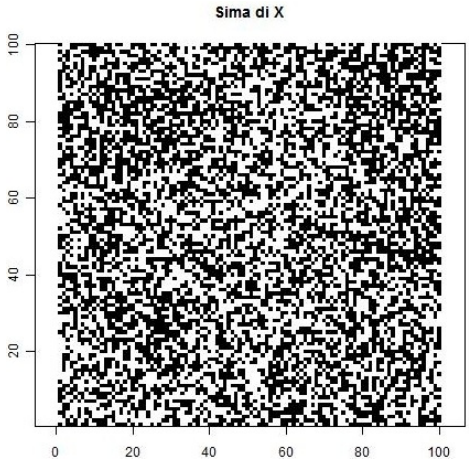
Figura 3.10: Prime quattro autofunzioni nel quinto caso, $\sigma^2 = 1$

In Figura 3.11 mostriamo i risultati ottenuti dalla classificazione non supervisionata dei dati del primo caso (quelli con i due cluster meno distinti). Le sei immagini, da sinistra a destra e dall'alto in basso, rappresentano il campo di Ising di partenza (lo stesso della Figura 3.1), il risultato della classificazione ottenuto applicando il metodo di spatial clustering con $\rho = 0.005$, 0.01, 0.02 e 0.025, e il risultato della classificazione dei dati originali (simple clustering), $\rho = 1$).

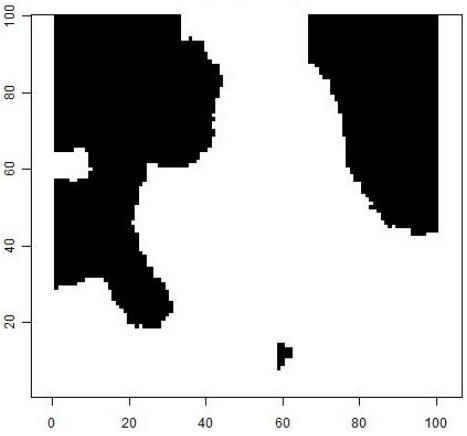
La Figura 3.11 conferma quanto avevamo affermato in maniera intuitiva: se i tasselli sono troppo piccoli (come negli ultimi due casi di Figura 3.11), la variabilità dei dati da classificare aumenta, rendendo la classificazione più difficile e peggiorando i risultati. In



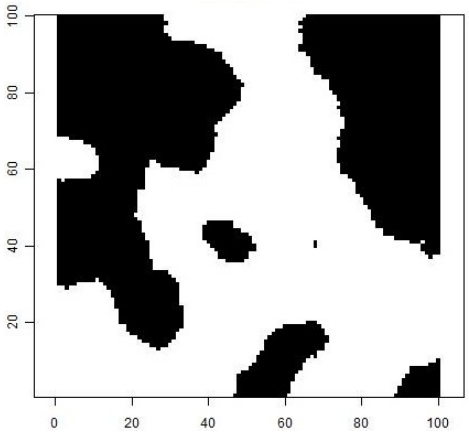
(a) Campo latente



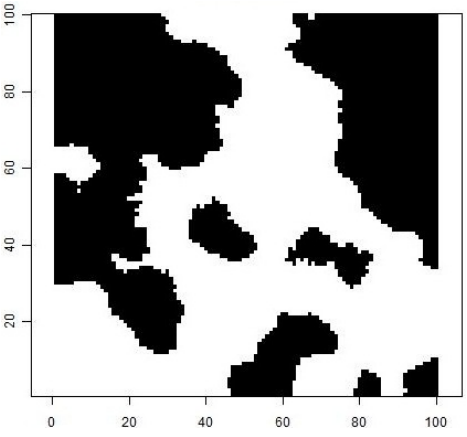
(b) Stima con simple clustering



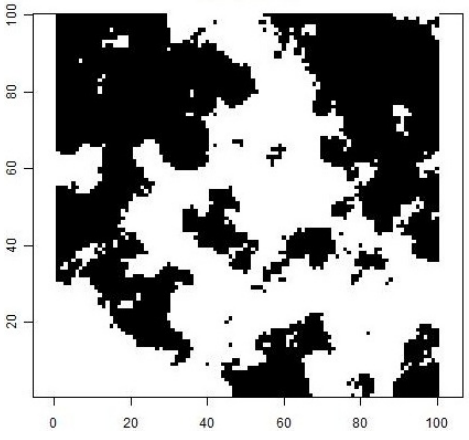
(c) Stima con spatial clustering $\rho = 0.005$



(d) Stima con spatial clustering $\rho = 0.01$



(e) Stima con spatial clustering $\rho = 0.02$



(f) Stima con spatial clustering $\rho = 0.025$

Figura 3.11: Nei primi due pannelli in alto: campo latente delle etichette (sinistra) e risultato del simple clustering (destra) sui dati del caso 1. Poi, in ordine orario partendo da (c), risultato dello spatial clustering sui dati del caso 1, rispettivamente per $\rho = 0.005$, 0.01, 0.02 e 0.025

questo caso vediamo chiaramente come, mentre la classificazione nel caso limite $\rho = 1$ non dà risultati apprezzabili (l'immagine originale si distingue appena), l'algoritmo proposto, per opportuni valori di ρ dà risultati decisamente migliori, anche con dati che provengono da due gruppi di etichette poco diversi tra loro. Osserviamo che in questo caso il risultato più somigliante al campo latente è quello ottenuto per $\rho = 0.01$ (al centro a destra), in quanto per valori più piccoli di ρ i tasselli sono troppo grandi (l'immagine risultante ha contorni molto più piatti di quella originale e non si riescono a identificare le zone più piccole), mentre per valori più grandi di ρ il risultato finale è un'immagine molto sgranata, a causa dell'alta variabilità dei dati di sintesi, che sono medie di tasselli molto piccoli.

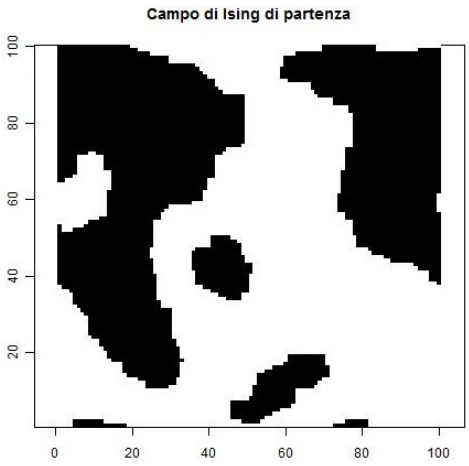
Infine, in Figura 3.12 vediamo gli stessi risultati relativi al secondo caso, nel quale i due gruppi sono più distinti. Innanzi tutto possiamo osservare che, come era prevedibile, per tutti i valori di ρ la classificazione è migliore che nel caso precedente. L'altro aspetto, meno scontato, che si può osservare dalle immagini è che la densità di tasselli rispetto alla quale la classificazione dà risultati migliori non dipende solo dall'immagine, ma anche dai dati. Infatti, mentre la classificazione per $\rho = 1$ è sempre la peggiore, questa volta otteniamo una stima migliore con un valore più elevato del parametro ($\rho = 0.025$), mentre nel caso precedente la classificazione migliore era quella ottenuta per $\rho = 0.01$. Questo avviene perché, se i dati dei due cluster sono tra loro molto diversi, non c'è bisogno di utilizzare anche le informazioni spaziali per classificarli. Si pensi per esempio ad una situazione nella quale i dati associati ai due gruppi siano nettamente distinti. Una classificazione dei dati originari porterebbe ad un errore molto basso, al limite uguale a zero, poiché la differenza tra i dati associati ai due gruppi è evidente. La tassellazione e classificazione dei dati medi invece aggiungerebbe comunque un errore dovuto alla distorsione del dataset, che è causata dalla presenza di tasselli che contengono dati di entrambi i gruppi. La stima ottenuta, in questo caso, sarebbe quindi meno precisa sui contorni dell'immagine, e l'errore si ridurrebbe solo diminuendo la distorsione introdotta nel dataset, ovvero diminuendo la taglia del tassello.

3.1.5 Studio del rate di misclassificazione al variare del parametro ρ

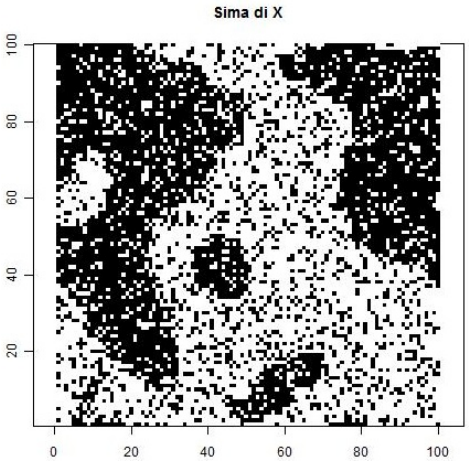
In base a tutte le considerazioni fatte precedentemente sulla taglia dei tasselli, se facciamo variare il parametro ρ tra 0 e 1 e calcoliamo ogni volta il rate di misclassificazione, ci aspettiamo di trovare un minimo per qualche valore di ρ. Ci aspettiamo inoltre che tale minimo sia più vicino a zero quanto più i dati originali provengono da gruppi vicini, e che sia in corrispondenza di $\rho = 1$ quando invece i due gruppi sono molto distinti.

Per verificare tale supposizione, utilizziamo l'algoritmo proposto per la classificazione di dati più o meno distinti (i 5 casi nelle Figure 3.2 e 3.3), in corrispondenza di diversi valori di ρ. Quello che otteniamo, rispettivamente per $\sigma^2 = 1$ e $\sigma^2 = 0.5$, sono i grafici nelle Figure 3.13 e 3.14. In particolare, i grafici mostrano l'andamento del rate di misclassificazione al variare del parametro ρ tra 0 e 0.025. Inoltre, la linea in rosso rappresenta il rate di misclassificazione nel caso di k–medie sui dati originali, quindi il valore per $\rho = 1$.

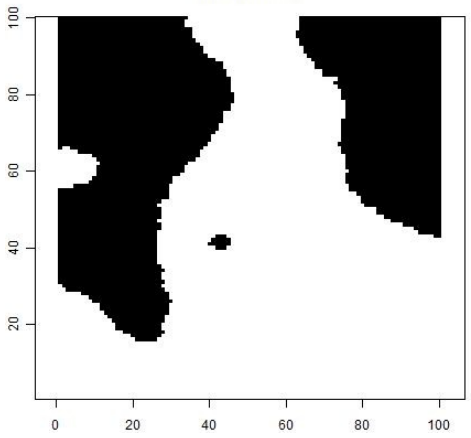
Quello che osserviamo giustifica molte delle precedenti supposizioni. Innanzi tutto c'è da osservare che l'algoritmo proposto dà risultati migliori delle k–medie in quasi tutti i casi: gli unici casi in cui la classificazione dei dati originali compie un errore minore rispetto all'algoritmo proposto per ogni valore di ρ analizzato, sono gli ultimi due con $\sigma^2 = 0.5$; si tratta cioè di quelle situazioni in cui i cluster originali sono talmente evidenti che non sono necessarie le informazioni spaziali. A giustificazione di questa affermazione, in Figura 3.15 presentiamo una frazione dei dati originali del settimo caso



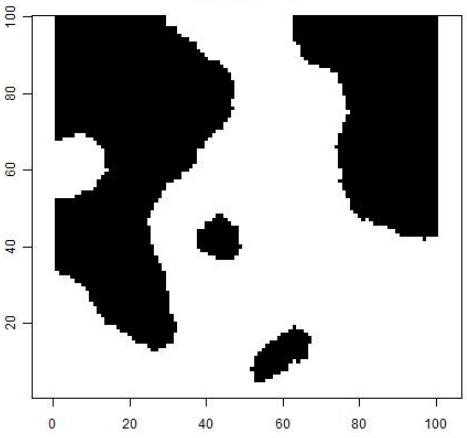
(a) Campo latente



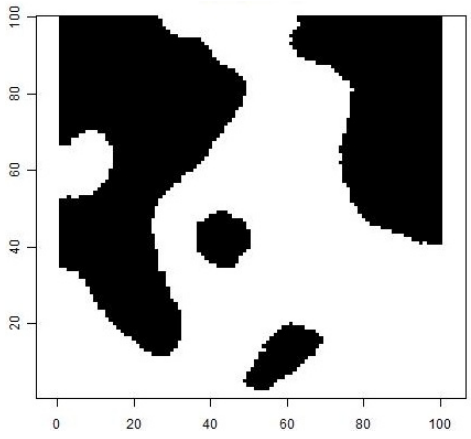
(b) Stima con simple clustering



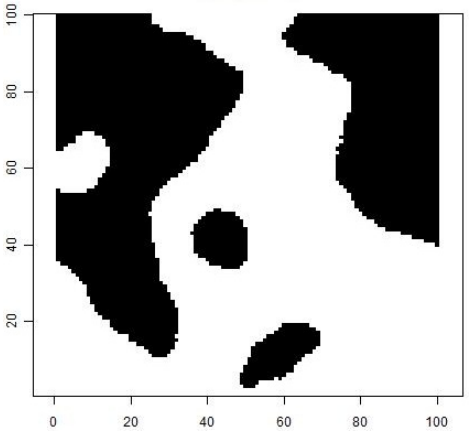
(c) Stima con spatial clustering $\rho = 0.005$



(d) Stima con spatial clustering $\rho = 0.01$



(e) Stima con spatial clustering $\rho = 0.02$



(f) Stima con spatial clustering $\rho = 0.025$

Figura 3.12: Nei primi due pannelli in alto: campo latente delle etichette (sinistra) e risultato del simple clustering (destra) sui dati del caso 5. Poi, in ordine orario partendo da (c), risultato dello spatial clustering sui dati del caso 5, rispettivamente per $\rho = 0.005$, 0.01, 0.02 e 0.025

di Figura 3.14 e i dati medi ottenuti dopo una tassellazione con $\rho = 0.025$: mentre osservando i dati originali, è abbastanza facile distinguere tra i dati associati ai due cluster (funzioni con tre massimi locali per un cluster e solo due per l'altro), i dati medi presentano il problema che già avevamo evidenziato all'inizio del capitolo, ovvero la transizione tra i dati associati ai due gruppi è graduale e non c'è una netta distinzione. Questo perché i dati medi associati a tasselli che ricoprono entrambi i cluster risultano essere una media tra funzioni di due gruppi diversi e quindi sono difficilmente associabili ad uno solo dei due gruppi, cosa che rende più difficile la classificazione, introducendo degli errori.

È più interessante notare, tuttavia, il gran numero di casi in cui l'algoritmo proposto si comporta molto meglio della classificazione dei dati originali. Osserviamo innanzi tutto che i casi appena discussi sono quelli più banali, dove la divisione nei due cluster è più facile da vedere. Quello che invece colpisce è che l'algoritmo che abbiamo analizzato riesce a suddividere i dati in due cluster con risultati abbastanza buoni anche nei casi peggiori, quelli cioè in cui i dati dei due gruppi sono più simili, come il caso analizzato nel Paragrafo 3.1.4 (che corrisponde al caso 1 nelle Figure 3.13 e 3.2).

Un'ulteriore osservazione può nascere considerando ancora una volta i risultati relativi ai casi in cui la classificazione dei dati è più difficile poiché le medie dei coefficienti di Fourier sono più vicine e la variabilità più alta, cioè i primi casi in Figura 3.13. Vediamo che l'andamento del rate di misclassificazione segue proprio l'andamento che avevamo supposto, presentando un minimo al variare delle dimensioni dei tasselli. Tale minimo, nei primi tre casi, si riesce ad osservare all'interno dei range considerati e si trova in corrispondenza di $\rho = 0.005$ nel caso 0, $\rho = 0.010$ nel caso 1 e $\rho = 0.020$ nel caso 2. Questo conferma la supposizione iniziale che il minimo rate di misclassificazione si sposta verso destra quanto più i dati di partenza sono divisi in due gruppi distinti. Nei casi 3-7 il minimo non si riesce ad osservare nei range considerati, ma supponendo che il rate di misclassificazione segua un andamento continuo al variare di ρ, tale minimo deve esistere poiché il valore per $\rho = 1$ è maggiore di quelli per valori intermedi di ρ.

Riassumendo quindi i risultati ottenuti da queste prime simulazioni, osserviamo che, al variare della diversità dei dati associati ai due gruppi:

- le prestazioni assolute del simple clustering e dello spatial clustering con $\rho = \rho_{OTT}$, cioè la densità di tasselli ottimale, migliorano sempre più i dati associati ai due gruppi sono diversi;
- in termini relativi, più i dati associati ai due gruppi sono simili, e quindi la classificazione è più difficile, più lo spatial clustering diventa vantaggioso rispetto al simple clustering, poiché l'informazione spaziale diventa sempre più importante ai fini della classificazione, rispetto a quella d'ampiezza.

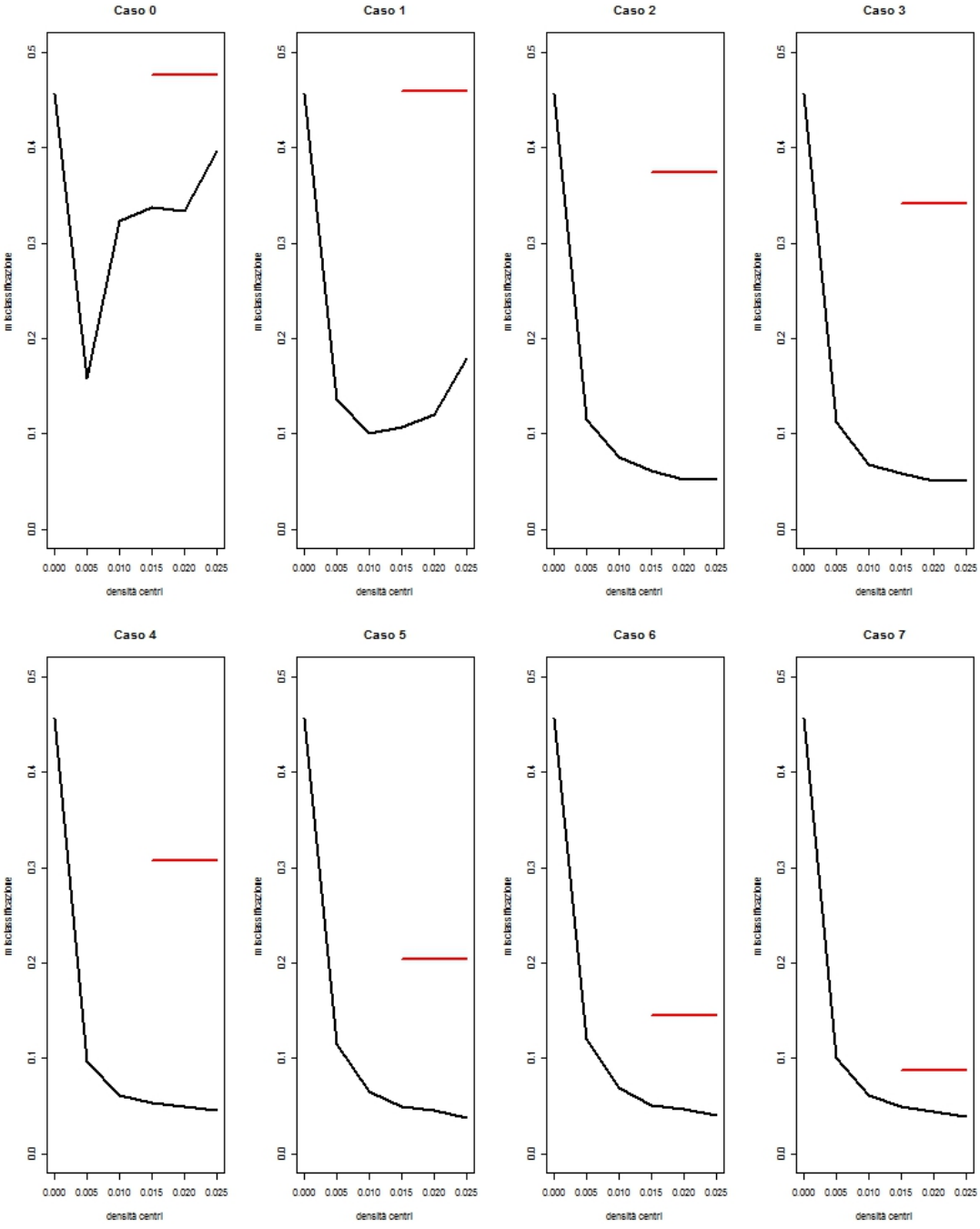


Figura 3.13: Rate di misclassificazione al variare di ρ, $\sigma^2 = 1$

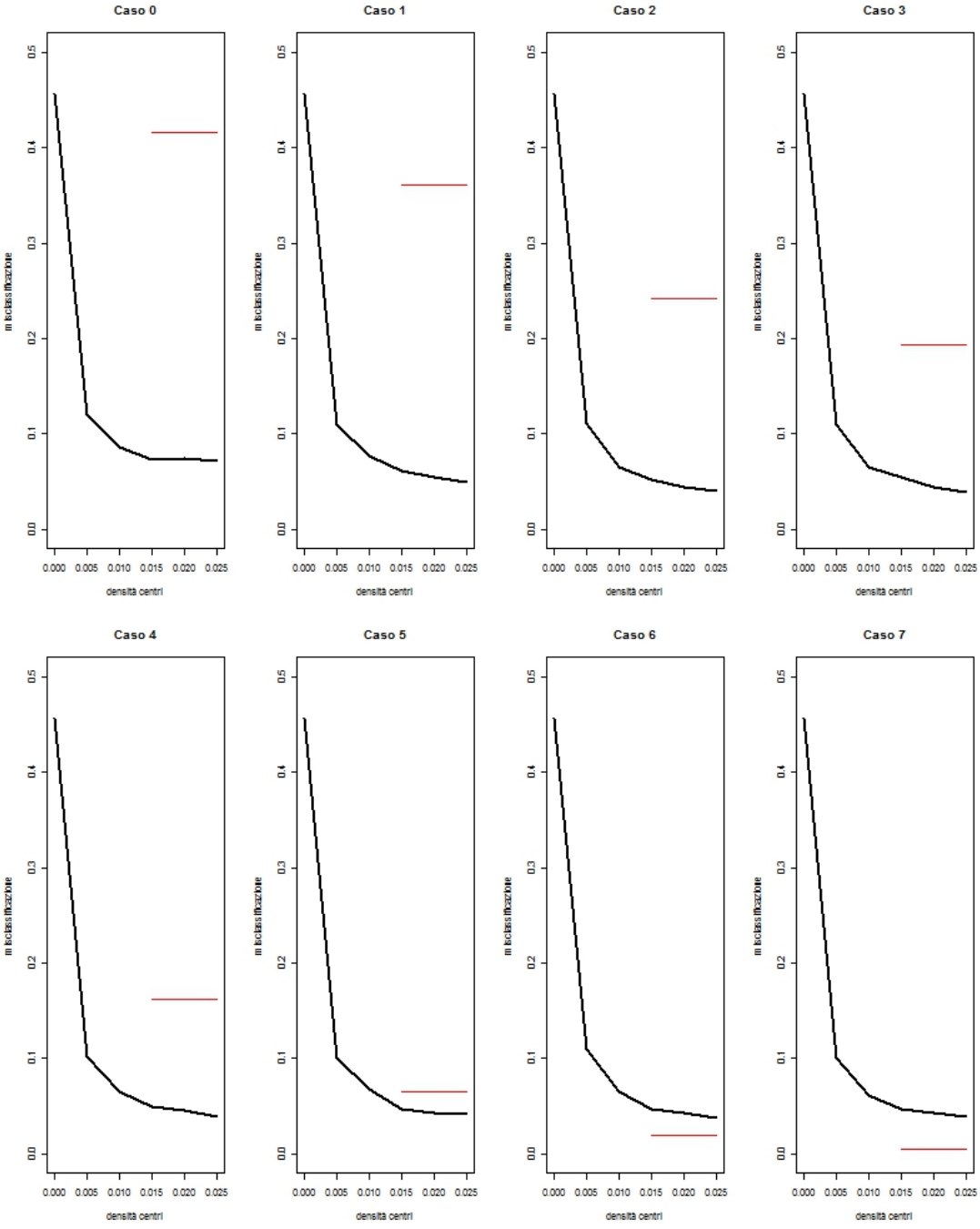
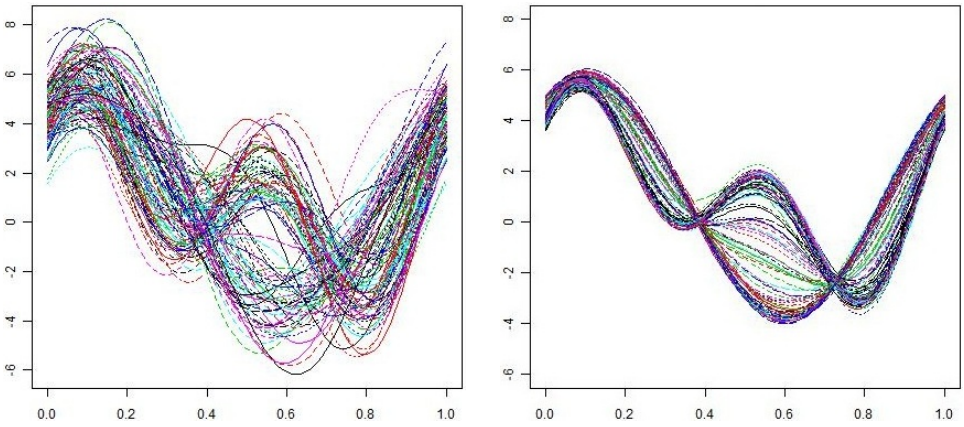


Figura 3.14: Rate di misclassificazione al variare di ρ, $\sigma^2 = 0.5$



(a) Campione casuale di 100 dati originali (b) Dati di sintesi con $\rho = 0.01$

Figura 3.15: Dati del settimo caso, $\sigma^2 = 0.5$

3.2 Simulazioni da un modello misto

Nella sezione precedente abbiamo analizzato le prestazioni dell'algoritmo descritto nel Capitolo 2 in presenza di un modello HMRF, la cui ipotesi principale è che i dati, condizionatamente alle etichette, siano indipendenti. Nella realtà tuttavia, è difficile pensare ad una situazione di totale indipendenza dei dati, una volta conosciute le etichette.

Possiamo dunque pensare ad un modello più generale, che si adatti meglio alla dipendenza spaziale effettivamente osservata nei dati reali dei casi applicativi, in cui tale dipendenza si esplica ad un duplice livello: una dipendenza spaziale nel campo latente delle etichette, che supponiamo influisca solo sulla media della distribuzione dei dati attraverso un modello di tipo HMRF, ed una dipendenza spaziale tra i dati stessi, che introduciamo tramite un'opportuna matrice di covarianza, Σ_{tot}, descritta da un covariogramma secondo le nozioni introdotte nella Sezione 1.2. Supponiamo, cioè, che la distribuzione dei vettori p–dimensionali dei coefficienti, condizionatamente alla realizzazione del campo latente delle etichette, sia una distribuzione normale multivariata la cui media dipende esclusivamente dalle etichette del campo latente, e la cui matrice di covarianza sia Σ_{tot}. Introducendo tale matrice, introduciamo quindi un secondo grado di dipendenza spaziale tra le realizzazioni del processo nei siti $\mathbf{s}_i$. Immaginiamo poi che questa seconda struttura di dipendenza esistente tra i vettori di coefficienti $\mathbf{c}_i$ associati agli N siti sia di tipo isotropo, che il processo sia stazionario e che la correlazione tra i dati decresca in maniera esponenziale con l'aumentare della distanza tra di essi. Siamo in presenza quindi di un modello misto, nel quale la dipendenza spaziale viene introdotta condizionatamente alle etichette a livello della media dei dati, e direttamente tra un dato e l'altro a livello della matrice di covarianza.

Osserviamo infine che il modello utilizzato precedentemente non è altro che un caso particolare dello schema ora presentato, nel quale la matrice di covarianza è diagonale.

3.2.1 Modello di generazione dei dati

Ipotizziamo, come nel caso precedente, di generare il dato funzionale simulando per ogni sito j un vettore p-variato di coefficienti della base di Fourier, che verrà costruito

come una realizzazione da una normale p-variata. Per simulare i dati dobbiamo quindi introdurre un'opportuna matrice di covarianza tra tutti i coefficienti di tutti i siti Σ_{tot}.

Innanzi tutto, ricordiamo che se la dimensione del campo delle etichette è $N_1 \times N_2$, con $N_1 = N_2 = \sqrt{N}$, e vogliamo simulare un dato funzionale in un sottospazio di dimensione p, avremo bisogno di una matrice di covarianza di dimensione $Np \times Np$. Tale matrice esprime la covarianza di un vettore che contiene tutti i coefficienti associati ai siti così costituito: allineiamo i siti del reticolo in un'unico vettore scorrendo il reticolo per righe, da sinistra a destra e dall'alto in basso. Una volta allineati i siti, ricordiamo che vogliamo simulare p coefficienti per ogni sito, quindi nel vettore finale ogni sito sarà associato a tre posizioni successive. A titolo di esempio, prendendo $p = 3$ e $\sqrt{N} = 5$, i primi tre elementi del vettore dei dati rappresentano le tre componenti associate al primo sito in alto a sinistra (il sito $(1, 1)$ prima riga, prima colonna), gli elementi da 4 a 6 le tre componenti associate al sito $(1, 2)$, e così via. Le tre componenti associate al sito per esempio, $(2, 1)$ sono quindi nelle posizioni 16, 17 e 18.

La matrice di covarianza del vettore di dati così costituito deve dunque avere una struttura a blocchi che tenga conto sia del tipo di dipendenza esistente tra i coefficienti associati ad uno stesso sito (covarianza del vettore $\mathbf{c_j}$), che di quella tra coefficienti associati a siti diversi. Definiamo quindi la matrice Σ_j di dimensione $p \times p$, come covarianza tra i coefficienti associati allo stesso sito $\mathbf{s}_j$. Supponiamo inoltre che la matrice Σ_j non dipenda dalle etichette del sito, né dal sito stesso: $\Sigma_j = \Sigma$ esattamente come nel caso precedente.

La diagonale della matrice Σ_{tot} a questo punto è costituita da una ripetizione del blocco Σ, mentre fuori dalla diagonale scegliamo di ripetere il medesimo blocco moltiplicato per un coefficiente che decresce in modo esponenziale con la distanza tra i due siti ai quali è associato il blocco. Stiamo quindi considerando un modello di covariogramma esponenziale come quello descritto nella Sezione 1.2, con una leggera modifica per considerare dati p-variati: i parametri del covariogramma sono ora delle matrici di dimensioni $p \times p$; in particolare il sill parziale è la matrice Σ, il nugget è anch'esso una matrice, che ipotizzeremo non dipendere dal sito j e che chiameremo Γ, mentre il parametro di decadimento $\Phi = \frac{1}{\lambda}$ (dove λ è il parametro di range) è reale. In conclusione, la matrice di covarianza $[\Sigma_{tot}]_{jj'}$ dei primi p coefficienti della base di Fourier tra due siti j e j' a distanza $d_{jj'}$ è data da:

$$[\Sigma_{tot}]_{jj'} = \begin{cases} \Sigma e^{-\frac{d_{jj'}}{\lambda}} & j \neq j' \\ \Sigma + \Gamma & j = j' \end{cases} \tag{3.1}$$

Per visualizzare l'effetto del coefficiente di decadimento esponenziale, mostriamo in Figura 3.16 una matrice ottenuta dall'equazione 3.1, con $\Gamma = 0$, Σ matrice costante, $\lambda = 1$, $p = 3$ e $\sqrt{N} = 5$. La Figura 3.16 è stata ottenuta da una matrice Σ costante esclusivamente per mostrare la struttura a blocchi del covariogramma risultante. Ovviamente, dovendo definire una valida matrice di covarianza, nelle simulazioni non utilizziamo una matrice costante, ma una reale matrice di covarianza che esprima la dipendenza tra i coefficienti della base di Fourier associati ad uno stesso sito. In generale, scegliamo matrici diagonali o matrici a dominanza diagonale stretta, per garantirne la definita positività, ottenendo le due strutture di dipendenza in Figura 3.17.

3.2.2 Piano di simulazione e scelta dei parametri

Per poter simulare dei dati dal modello descritto, abbiamo bisogno di poter memorizzare e fattorizzare una matrice di covarianza di grandi dimensioni, cioè la matrice Σ_{tot} sopra

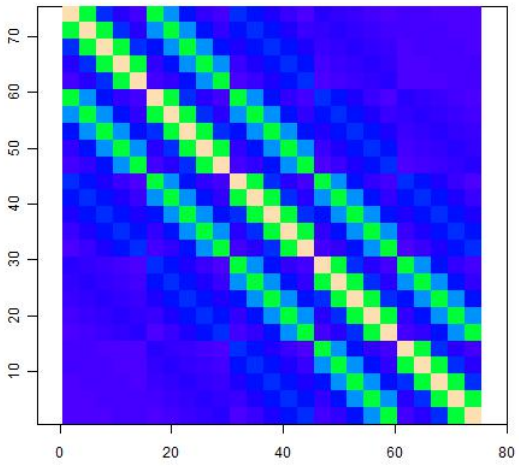


Figura 3.16: Struttura di decadimento esponenziale nei blocchi della matrice di covarianza per $\sqrt{N} = 5$, $p = 3$

descritta. Per questo motivo, per le simulazioni è necessario ridurre la dimensione del campo di Ising di partenza ed il numero di coefficienti della base di Fourier considerati. Scegliamo quindi $N = 50$ e $p = 3$ (la matrice Σ_{tot} ha così dimensioni 7500×7500).

Come nel caso precedente, per la distribuzione spaziale delle etichette utilizziamo una realizzazione di un campo di Ising con $\beta = 2$. La medesima realizzazione sarà utilizzata per tutte le simulazioni ed è quella mostrata in Figura 3.18.

Simuliamo i dati da tre modelli diversi:

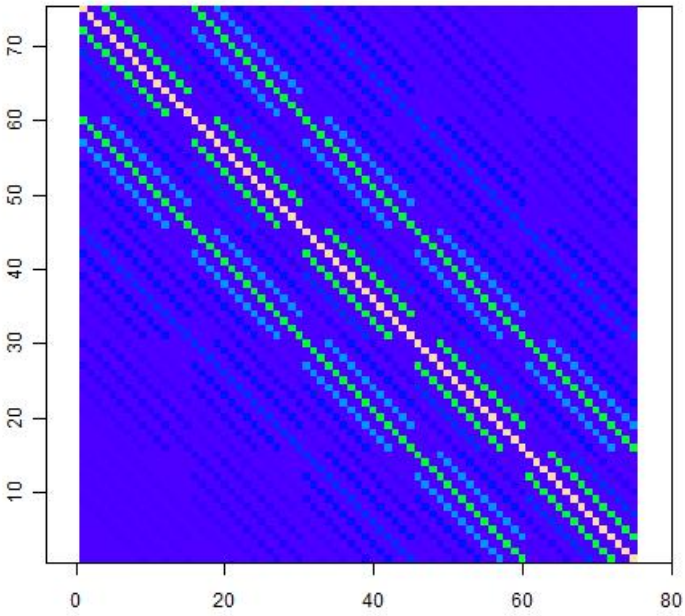
1. modello senza nugget e matrice di covarianza dei coefficienti diagonale: $\Gamma = 0$, $\Sigma = I$;

2. modello senza nugget e matrice di covarianza dei coefficienti a dominanza diagonale stretta: $\Gamma = 0$, $\Sigma = \Sigma_1 = \begin{pmatrix} 1.3 & 0.5 & 0.1 \\ 0.5 & 1.3 & 0.5 \\ 0.1 & 0.5 & 1.3 \end{pmatrix}$;

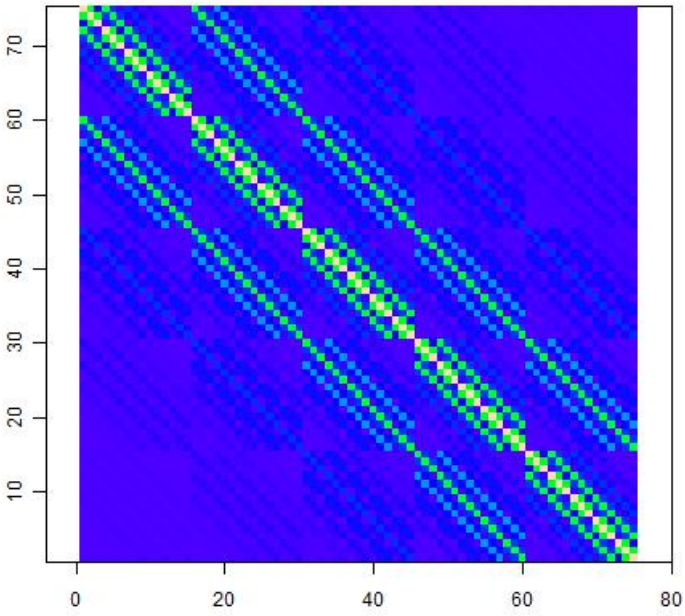
3. modello con nugget e matrice di covarianza dei coefficienti diagonale: $\Gamma = \Sigma = I$.

Per la scelta del parametro di range λ, analizziamo per ogni modello i risultati relativi a tre diversi valori del parametro: $\lambda = 0$, che rappresenta il caso in cui la dipendenza spaziale è condizionata solamente alle etichette, e dati in siti diversi sono indipendenti condizionatamente alle etichette, $\lambda = 0.5$ e $\lambda = 1$. Facendo variare il parametro λ, vogliamo osservare l'effetto sulla classificazione dell'aggiunta di un altro grado di dipendenza spaziale tra i dati. All'aumentare di questo parametro, la dipendenza spaziale tra i dati aumenta, e il modello si differenzia dal caso analizzato nella Sezione 3.1, che corrisponde a $\lambda = 0$.

Per quanto riguarda i vettori delle medie dei coefficienti dei due gruppi, scegliamo come nel caso precedente di farli variare considerando otto casi, partendo dai primi in cui le medie dei gruppi sono quasi coincidenti, per arrivare agli ultimi nei quali i due gruppi di dati sono tra loro molto distinti. I coefficienti utilizzati sono riassunti in Tabella 3.2, mentre i dati funzionali ottenuti dai coefficienti medi e dai valori medi ai



(a) Σ diagonale



(b) Σ a dominanza diagonale stretta

Figura 3.17: Esempi di struttura della matrice di covarianza del vettore dei dati per $\sqrt{N} = 5$, $p = 3$

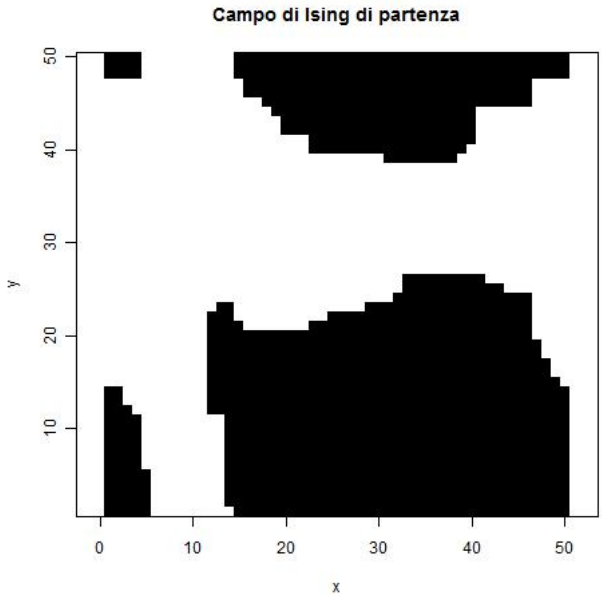


Figura 3.18: Realizzazione di un campo di Ising con $\beta = 2$

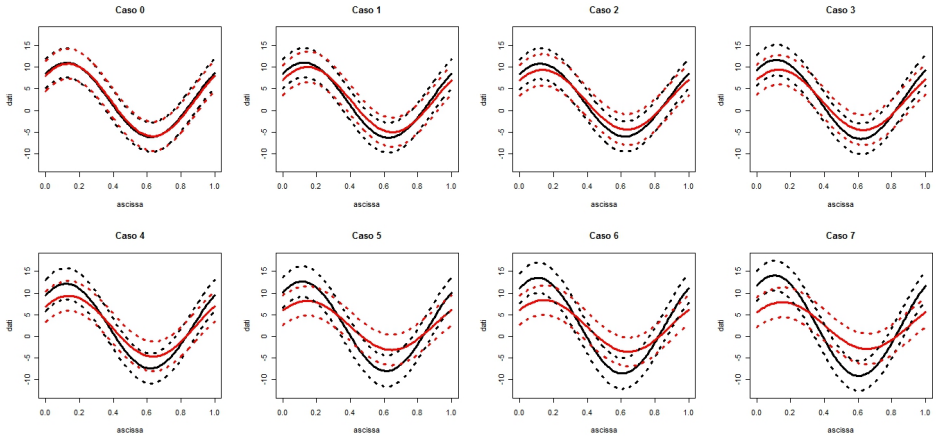
quali abbiamo sommato o sottratto due volte la deviazione standard puntuale per i tre casi sono presentati in Figura 3.19. Osserviamo dalla Figura 3.19 che la variabilità dei

	$(\mu_{-1})_1$	$(\mu_{-1})_2$	$(\mu_{-1})_3$	$(\mu_1)_1$	$(\mu_1)_2$	$(\mu_1)_3$
Caso 0	2.415	4.308	4.231	2.407	4.428	3.896
Caso 1	2.448	4.410	4.258	2.495	4.319	3.172
Caso 2	2.318	4.212	4.273	2.486	3.730	3.166
Caso 3	2.447	4.309	4.742	2.425	3.667	3.298
Caso 4	2.342	4.877	4.945	2.425	3.667	3.298
Caso 5	2.376	4.830	5.438	2.545	3.147	2.549
Caso 6	2.520	4.915	6.003	2.446	3.328	2.601
Caso 7	2.455	4.930	6.443	2.512	3.170	2.132

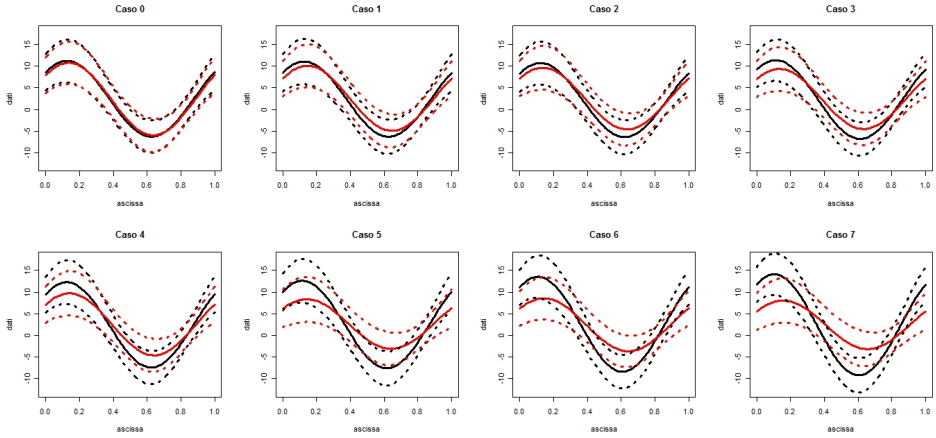
Tabella 3.2: Medie delle distribuzioni dei coefficienti associate alle due etichette nel modello misto

dati è leggermente maggiore negli ultimi due modelli, cosa che ci potevamo aspettare osservando la matrice Σ.

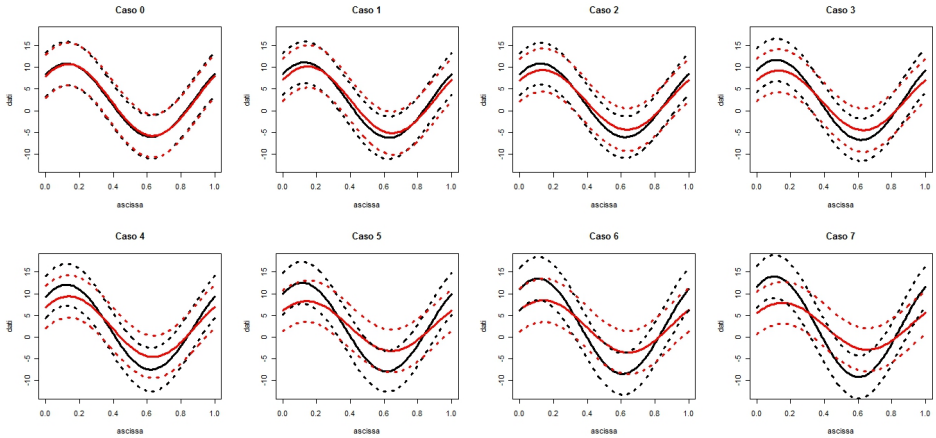
Per quanto riguarda, infine, il parametro ρ, manteniamo lo schema proposto nella prima parte del capitolo: faremo variare la densità dei tasselli per poter osservare se esiste un minimo dell'errore di misclassificazione al variare di ρ, e in caso affermativo cercheremo di capire la dipendenza di tale minimo dai dati di origine.



(a) Modello 1: $\Sigma = I$, $\Gamma = 0$



(b) Modello 2: $\Sigma = \Sigma_1$, $\Gamma = 0$



(c) Modello 3: $\Sigma = \Gamma = I$

Figura 3.19: Dati funzionali ottenuti con i coefficienti dati dai valori medi $(-)$ e dai valori medi puntuali $\pm$ due volte la deviazione standard puntuale per i tre modelli

3.2.3 Risultati delle simulazioni

Vogliamo ora analizzare i risultati ottenuti applicando i metodi di classificazione utilizzati nel piano di simulazione della Sezione 3.1 (simple clustering, spatial clustering e trivial clustering) al modello misto appena descritto. I risultati ottenuti per diversi valori del parametro ρ per i tre diversi modelli sono presentati nelle Figure 3.20, 3.21 e 3.22. I grafici presentano i risultati da noi ottenuti per ogni modello descritto per diversi valori di λ: la linea rossa corrisponde al caso di indipendenza condizionata alle etichette ($\lambda = 0$), la linea verde corrisponde al caso $\lambda = 0.5$, mentre quella blu corrisponde a $\lambda = 1$.

Per prima cosa osserviamo che, come per il modello nel quale la dipendenza spaziale è condizionata solamente alle etichette, nella maggior parte dei casi, i risultati della classificazione con l'algoritmo proposto da risultati nettamente migliori sia della classificazione banale, che delle k–medie su tutti i dati, soprattutto nei casi in cui i cluster sono meno distinti. Si riesce inoltre, nei primi casi, ad osservare il minimo in funzione della densità dei tasselli. Anche per il modello misto, valgono le stesse considerazioni fatte per il modello di indipendenza condizionatamente alle etichette, cioè il fatto che esiste un valore di ρ intermedio per il quale l'errore è minimo, e tale valore si sposta verso destra al crescere della distinzione tra i due cluster. Nei casi in cui i cluster sono molto distinti il minimo coinciderà con $\rho = 1$, cioè la classificazione di tutti i dati trascurando l'ipotesi di dipendenza spaziale.

Per quanto riguarda il comportamento del rate di misclassificazione al variare del parametro λ, e quindi all'aumentare del grado di dipendenza spaziale introdotto tramite il covariogramma, i risultati non sono così interpretabili. Dal momento che aumentando il valore di λ il grado di dipendenza spaziale aumenta, ci si aspetterebbe che i risultati dell'algoritmo proposto, che si basa sull'ipotesi che ci sia dipendenza tra i dati, migliorino, poiché per valori di λ alti la dipendenza spaziale è maggiore. Tuttavia, osservando i risultati ottenuti dalla classificazione dei dati dei tre modelli, vediamo che questo non succede. Al contrario, aumentando λ i risultati peggiorano leggermente, soprattutto nei primi casi, quelli in cui, cioè, la differenza tra i vettori delle medie è molto piccola.

Questo comportamento deriva dal fatto che, proprio perché con l'introduzione della matrice Σ_{tot} abbiamo aumentato il grado di dipendenza tra dati associati a diversi siti, abbiamo introdotto una dipendenza tra i siti che agisce direttamente sulla realizzazione di Y nel sito, modificandone la variabilità, e dunque rendendo più difficile la stima della media, che è legata alle etichette del campo latente oggetto del metodo di classificazione. L'algoritmo proposto, classifica i dati effettuando la classificazione sul dataset dei dati di sintesi, che costituiscono stime delle medie locali del processo Y. Quindi, probabilmente, tale algoritmo dà risultati migliori nel caso in cui la dipendenza spaziale tra i siti si presenti esclusivamente tramite un modello HMRF, nel quale viene introdotta una differenza nelle medie delle distribuzioni di emissione associate alle diverse etichette, e la matrice di covarianza associata ad ogni sito è costante. Nel caso in cui, invece, la dipendenza spaziale è introdotta con modelli diversi da un HMRF, come per esempio il modello misto proposto, la classificazione tramite spatial clustering è meno affidabile, proprio perché non considera la struttura di covarianza spaziale.

In particolare, nel caso dei dati generati con il modello misto, osserviamo che nei casi in cui le medie dei coefficienti associati ai due gruppi sono tra loro poco distinte, e quindi la dipendenza spaziale tra i siti si manifesta maggiormente tramite la struttura del covariogramma, abbiamo nella classificazione degli errori sistematici, che creano una sorta di distorsione dell'immagine di partenza, indipendentemente dal numero di tasselli utilizzati.

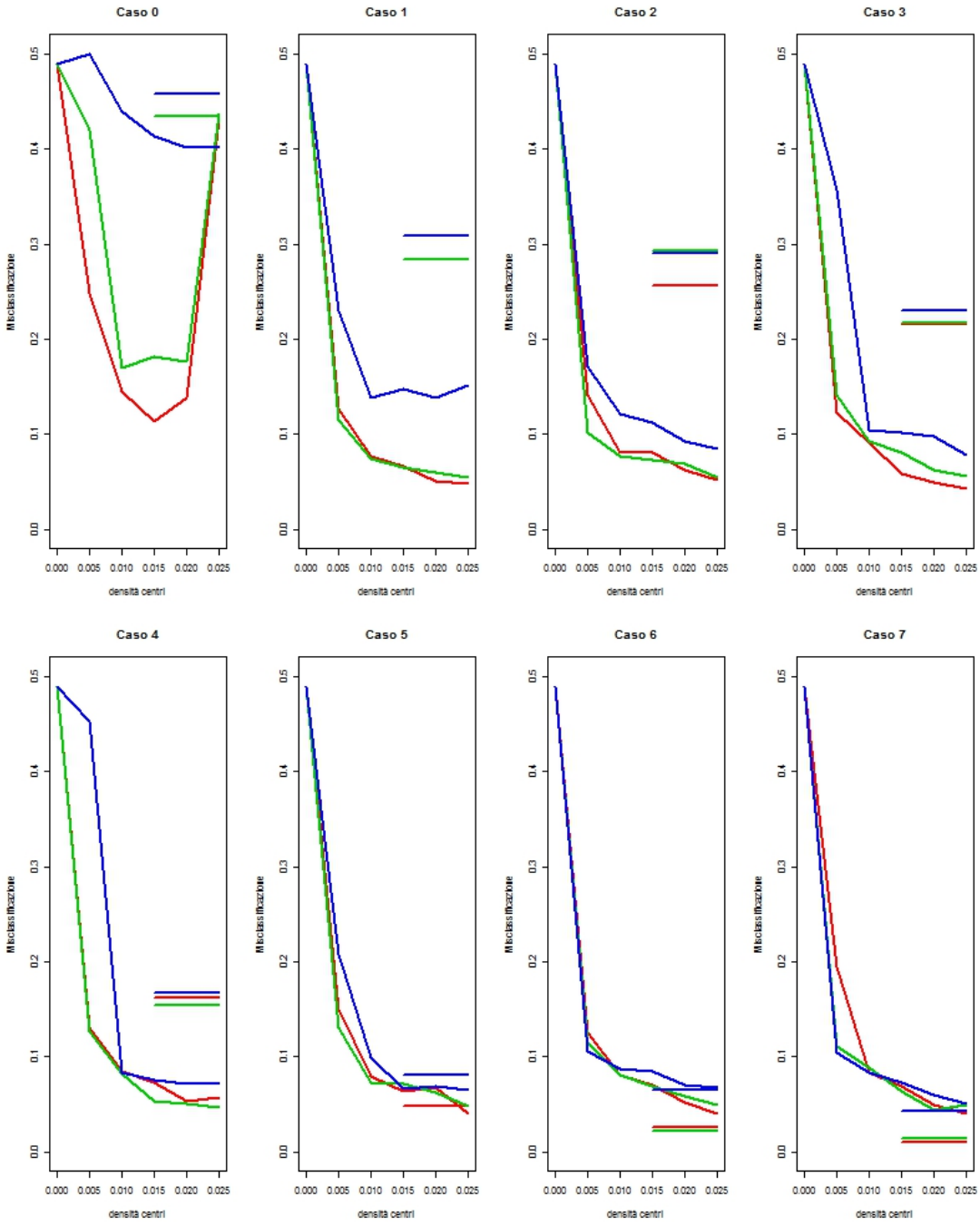


Figura 3.20: Rate di misclassificazione al variare di ρ per il primo modello ($\Sigma = I$, $\Gamma = 0$) per $\lambda = 0$ (linea rossa), $\lambda = 0.5$ (linea verde) e $\lambda = 1$ (linea blu)

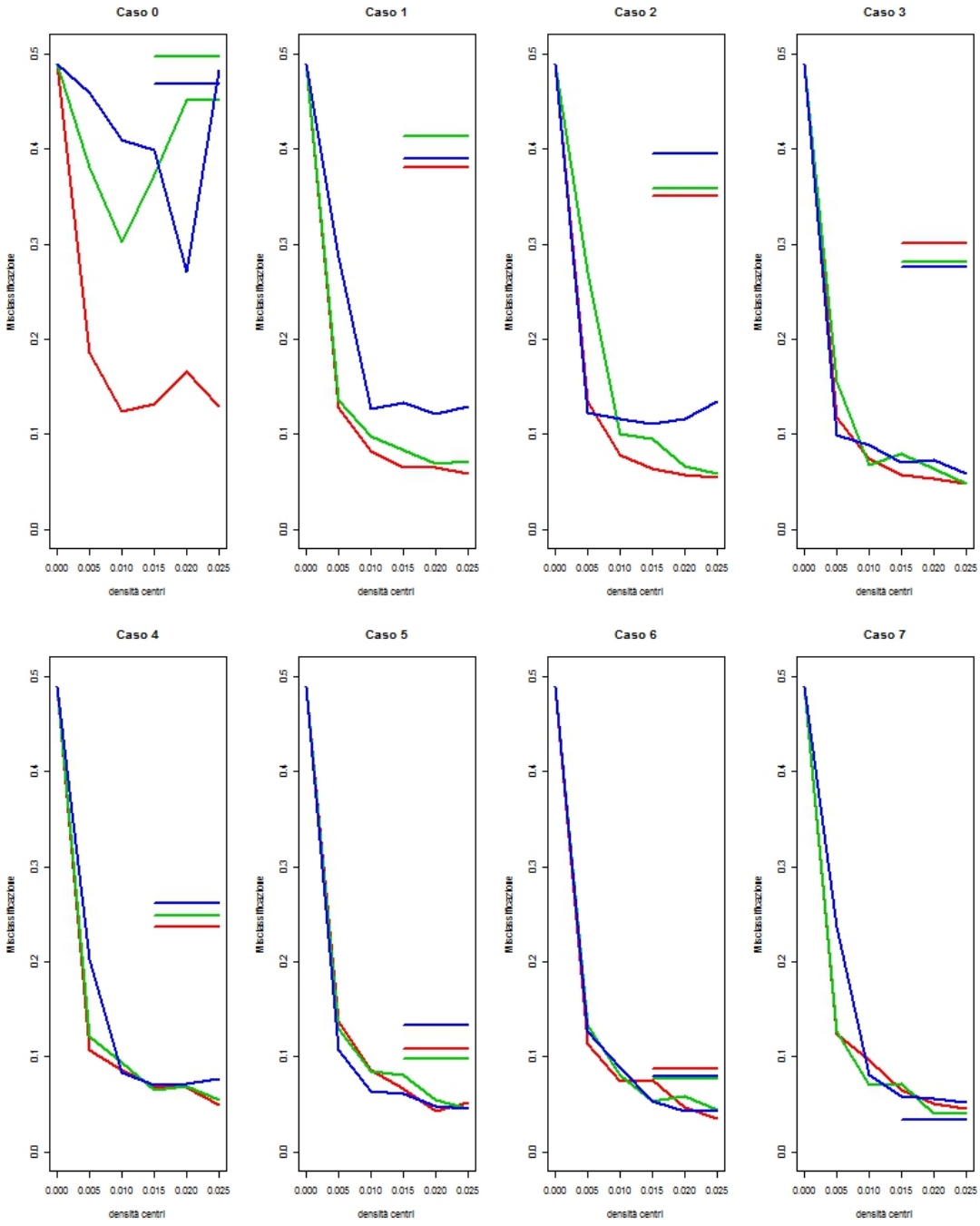


Figura 3.21: Rate di misclassificazione al variare di ρ per il secondo modello ($\Sigma = \sigma_1$, $\Gamma = 0$) per $\lambda = 0$ (linea rossa), $\lambda = 0.5$ (linea verde) e $\lambda = 1$ (linea blu)

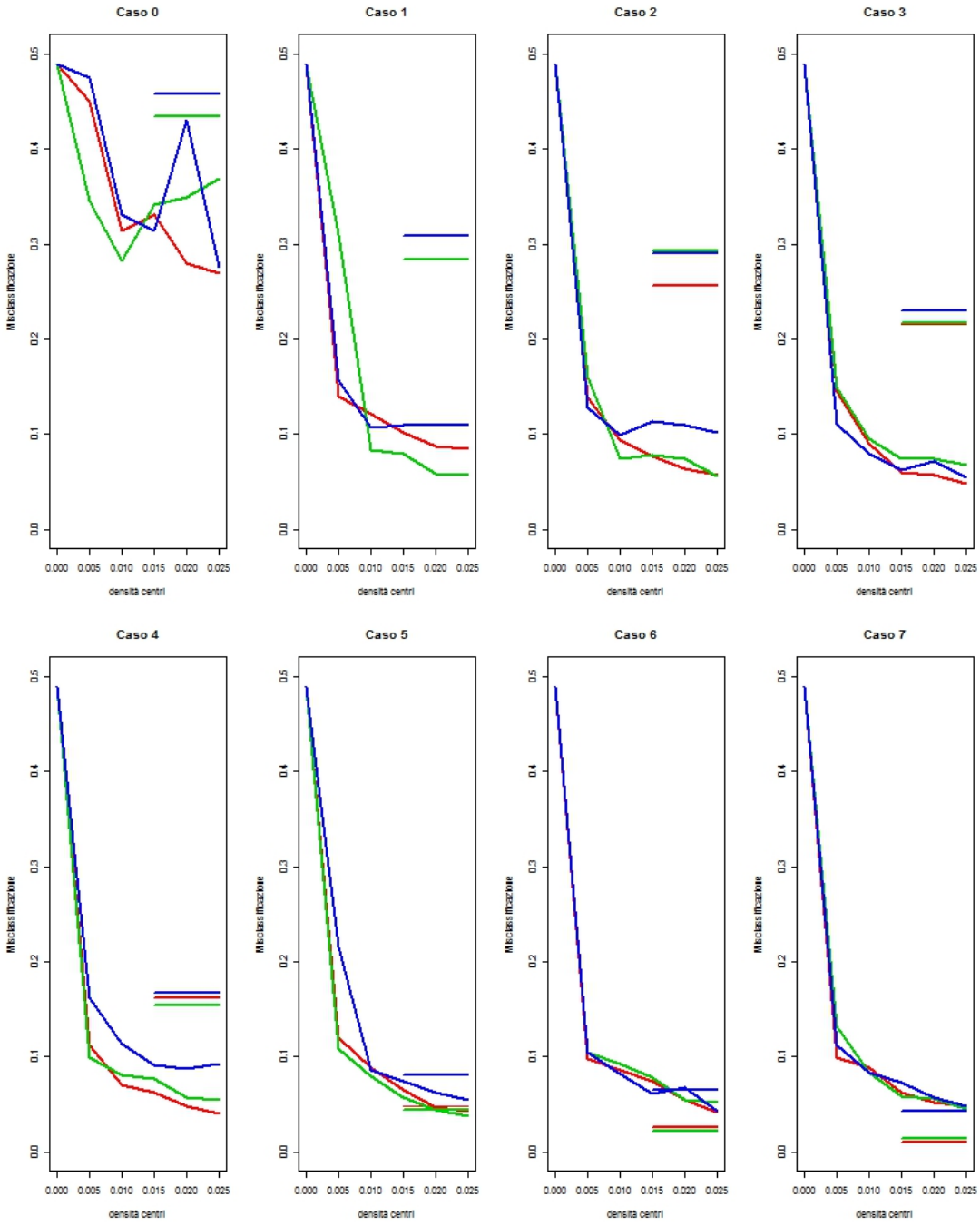
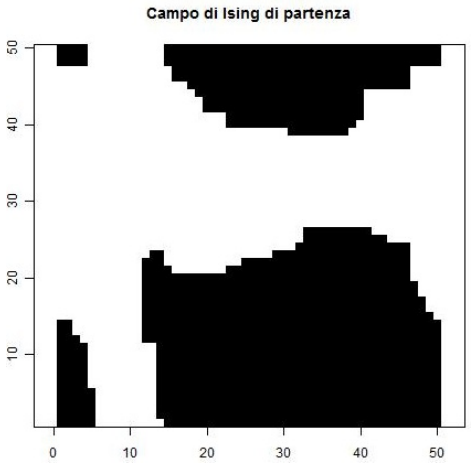
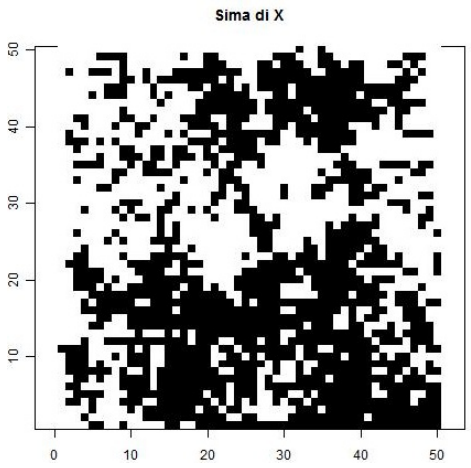


Figura 3.22: Rate di misclassificazione al variare di ρ per il terzo modello ($\Sigma = \Gamma = I$) per $\lambda = 0$ (linea rossa), $\lambda = 0.5$ (linea verde) e $\lambda = 1$ (linea blu)

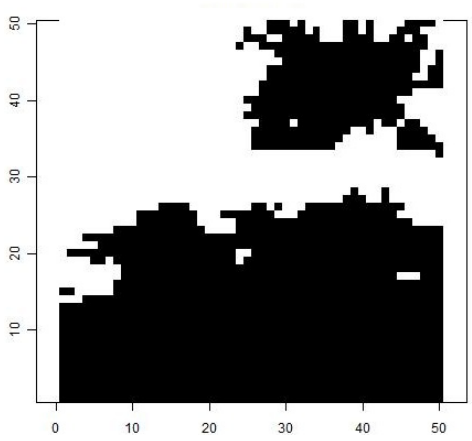
Per osservare l'effetto appena descritto, presentiamo in Figura 3.23 i risultati ottenuti classificando i dati del primo modello (con matrice Σ diagonale e senza nugget), con coefficienti medi nei due cluster molto vicini e dati da $\boldsymbol{\mu}_{-1} = (2.448, 4.410, 4.258)$ e $\boldsymbol{\mu}_1 = (2.495, 4.319, 3.172)$, e $\lambda = 1$ (quindi la seconda dipendenza spaziale introdotta col covariogramma è più forte). Le etichette del campo latente che vogliamo ricostruire sono raffigurate nel pannello in alto a sinistra in Figura 3.23. Notiamo che i risultati della classificazione si discostano in maniera sistematica dalle etichette originali: i contorni delle classificazioni ottenute sono leggermente diversi, e si discostano dal campo latente sempre nello stesso modo.



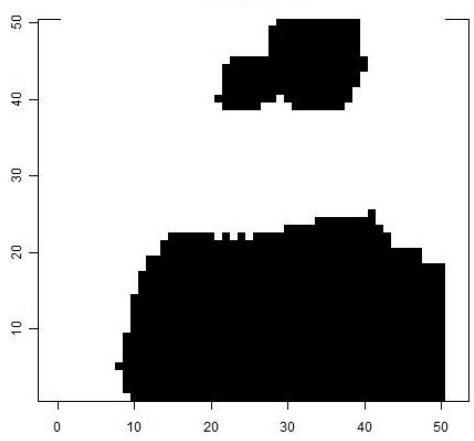
(a) Campo latente



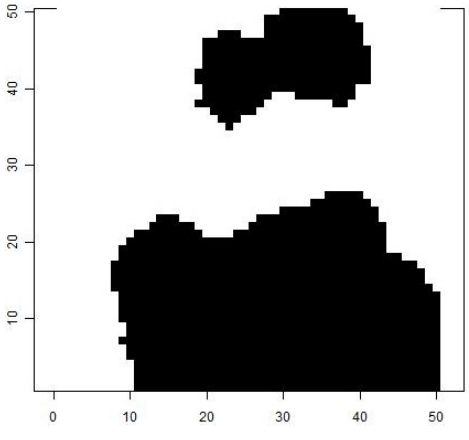
(b) Stima con simple clustering



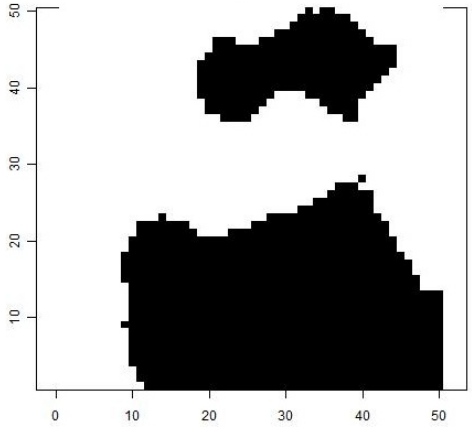
(c) Stima con spatial clustering $\rho = 0.005$



(d) Stima con spatial clustering $\rho = 0.01$



(e) Stima con spatial clustering $\rho = 0.02$



(f) Stima con spatial clustering $\rho = 0.025$

Figura 3.23: Nei primi due pannelli in alto: campo latente delle etichette (sinistra) e risultato del simple clustering (destra) sui dati del caso 1 con parametri $\Sigma = I, \Gamma = 0$ e $\lambda = 1$. Poi, in ordine orario partendo da (c), risultato dello spatial clustering sui dati del caso 5, rispettivamente per $\rho = 0.005$, 0.01, 0.02 e 0.025

Capitolo 4

Analisi di dati climatici

4.1 Meteorologia per l'energia solare

Nel corso di questo capitolo, mostreremo come sia possibile declinare le tecniche descritte nel Capitolo 2 per l'analisi di dati spazialmente dipendenti provenienti da un'applicazione reale in ambito climatico. In particolare, scegliamo di analizzare dati climatici georeferenziati, associati all'irraggiamento solare mensile in ciascun luogo del pianeta.

I dati che analizzeremo provengono dalla release 6 (2008) del database *Surface meteorology and Solar Energy* (SSE) della NASA [NASA, 2008]. Il database SSE si situa all'interno del progetto POWER (Prediction Of Worldwide Energy Resource) della NASA, e si pone l'obiettivo di facilitare l'utilizzo in tutto il mondo di energie rinnovabili. Per questo, raccoglie moltissime informazioni legate all'utilizzo dell'energia solare, o ad altre energie rinnovabili in una qualsiasi zona della terra, al fine di fornire tutti i parametri necessari per la progettazione e costruzione di impianti a energia rinnovabile, come per esempio impianti fotovoltaici, o impianti a energia eolica.

In particolare, all'interno di tale database, sono disponibili informazioni giornaliere, mensili o annuali riguardanti insolazione, temperatura, direzione e forza del vento, per ciascuna coordinata geografica. Inoltre, molto più utile dal nostro punto di vista, sono disponibili dataset globali riguardanti alcuni parametri di interesse, contenenti misurazioni medie mensili di un dato parametro registrate sulle celle di una griglia formata da meridiani e paralleli in cui ogni elemento è un 'quadrato curvo' di lato $1°$.

Un dataset globale contiene quindi 12 misurazioni mensili di una grandezza Y per ogni area $A_{\lambda,\theta} = [\lambda, \lambda+1] \times [\theta, \theta+1]$, dove λ e θ sono rispettivamente due valori interi di longitudine e latitudine: $\lambda \in \mathbb{Z} \cap [-180, 179], \quad \theta \in \mathbb{Z} \cap [-90, 89]$; valori negativi di latitudine e longitudine rappresentano, rispettivamente, l'emisfero australe e la metà del globo a ovest del meridiano di Greenwhich e a est della linea del cambiamento di data (emisfero occidentale).

Per la nostra analisi, ci concentreremo sull'energia solare, quella per la quale il database SSE contiene maggiori informazioni. La prima grandezza di riferimento per la progettazione di un impianto fotovoltaico, è l'insolazione. L'insolazione è la misura della quantità di radiazione emessa dal sole che raggiunge una data superficie in ogni unità di tempo, espressa in kilowattora per metroquadro al giorno $kwh/(m^2 day)$. La superficie, nel caso del database SSE, può essere semplicemente una superficie piana posta orizzontalmente sulla superficie terrestre, oppure una superficie piana ruotata, che rappresenta il caso più interessante dal punto di vista operativo.

La misura dell'insolazione media mensile, per una data località geografica, è indispensabile per la progettazione di impianti fotovoltaici, poiché rappresenta una misura

della quantità di energia disponibile all'ora per un metro quadrato di pannello.

Tale grandezza è stata dedotta a partire da osservazioni satellitari, nell'ambito di diversi progetti NASA, per un periodo di 22 anni, da luglio 1983 a giugno 2005. Non abbiamo a disposizione i dati mensili anno per anno, ma esclusivamente la stima dell'insolazione media mensile sul periodo di riferimento di 22 anni per ogni area $A_{\lambda,\theta}$.

4.1.1 La variabile di interesse

L'insolazione incidente sulla superficie terrestre è ovviamente un parametro molto importante se si vuole progettare un impianto fotovoltaico in una data regione. Tuttavia, diversamente da quanto si può pensare, tale parametro non è l'unico da prendere in considerazione qualora si voglia operare un confronto tra diverse regioni a livello globale, con l'obiettivo di ottimizzare la scelta dell'ubicazione di un impianto fotovoltaico.

Nella nostra analisi non considereremo l'insolazione, bensì una quantità da essa derivata, ovvero il numero equivalente di giorni consecutivi senza sole. Tale variabile è molto importante per il dimensionamento di un eventuale impianto a celle fotovoltaiche: infatti, in generale, si vuole che l'erogazione di energia elettrica sia continua durante il periodo di funzionamento dell'impianto, sia di giorno che di notte, ovvero che non ci siano cioè periodi nei quali tale erogazione sia nulla o molto scarsa. Tuttavia, l'insolazione incidente non è mai continua, a causa di condizioni meteorologiche variabili, o anche solo a causa dell'alternanza tra giorno e notte. Per questo motivo è essenziale che un impianto fotovoltaico non alimentato da una rete elettrica esterna sia collegato ad una batteria o ad un qualsiasi sistema di backup in grado di immagazzinare energia nei periodi di elevata insolazione, per poi rilasciarla nei periodi in cui l'insolazione è scarsa o nulla, in modo da erogare in maniera costante una quantità di energia pari all'insolazione media su un determinato periodo, per esempio su un mese.

Tali batterie o sistemi di backup devono quindi essere scelti in modo da immagazzinare una quantità di energia sufficiente ad alimentare l'impianto in ogni situazione. Per questo, un parametro molto importante per il dimensionamento di tali sistemi è il numero massimo di giorni equivalenti consecutivi senza sole per un dato mese, cioè il numero massimo di giorni durante i quali, in un dato mese, il sistema di backup deve essere in grado di fornire la stessa quantità di energia media mensile senza possibilità di ricaricarsi.

In prima analisi, un approccio conservativo potrebbe essere scegliere il massimo lungo l'intero arco dell'anno della quantità in questione, o anche il massimo negli ultimi 22 anni, per dimensionare i sistemi di backup in modo che gli impianti funzionino anche nel peggiore dei casi. Non sembrerebbe quindi necessario prendere in considerazione l'andamento annuale di tale parametro, ovvero considerare il dato come funzionale.

Potremmo però essere interessati alla progettazione di impianti che debbano funzionare esclusivamente in determinati periodi dell'anno (si pensi ai condizionatori d'estate o i sistemi di riscaldamento d'inverno). Inoltre, considerare esclusivamente il dato corrispondente al numero massimo di giorni consecutivi senza sole non permette in alcun modo di studiarne l'andamento annuale, al fine di identificare zone all'interno delle quali l'evoluzione annuale dei periodi di copertura nuvolosa è simile.

Per un'analisi più approfondita, quindi, scegliamo di prendere in considerazione l'andamento annuale di tale parametro, considerando i dati a disposizione come realizzazioni mensili di una funzione aleatoria con dominio annuale, per poi utilizzare le tecniche proposte per l'analisi e classificazione di dati funzionali.

4.1.2 Calcolo del numero massimo di giorni equivalenti consecutivi senza sole per un dato mese

Prima di effettuare le analisi di classificazione, presentiamo in dettaglio le caratteristiche del dataset a nostra disposizione. Come per l'insolazione, si tratta di dati mensili relativi al periodo compreso tra luglio 1983 e giugno 2005. Anche in questo caso si tratta di dati medi su una regione di un grado di latitudine per un grado di longitudine: abbiamo quindi una griglia di 180×360 siti, per ognuno dei quali abbiamo a disposizione 12 dati, relativi ai mesi dell'anno.

Nel paragrafo precedente abbiamo accennato al fatto che il numero di giorni senza sole non è una grandezza misurata per ogni sito, ma derivata dal valore dell'insolazione media mensile nel sito stesso: vediamo ora esplicitamente come tale grandezza è ricavata, a partire dai valori dell'insolazione media mensile, e dell'insolazione minima disponibile, sullo stesso periodo di tempo.

Immaginiamo, per semplicità, di osservare per 4 anni consecutivi i valori dell'insolazione giornaliera (in $kwh/m^2/giorno$) durante un intero mese (che supponiamo essere di 30 giorni) in una particolare località. Supponiamo, in un caso limite didattico, di avere osservato nei quattro anni un'insolazione costante, fatta eccezione per tre giorni consecutivi del primo anno, nei quali l'insolazione incidente è stata nulla. I dati osservati nel caso descritto, la loro media giornaliera e l'insolazione minima sono riassunti in Tabella 4.1.

anno-giorno	1	2	3	4	5	6	7	8	$\cdots$	30
1	2	2	0	0	0	2	2	2	$\cdots$	2
2	2	2	2	2	2	2	2	2	$\cdots$	2
3	2	2	2	2	2	2	2	2	$\cdots$	2
4	2	2	2	2	2	2	2	2	$\cdots$	2
media	2	2	0.75	0.75	0.75	2	2	2	$\cdots$	2
minimo	2	2	0	0	0	2	2	2	$\cdots$	2

Tabella 4.1: Dati relativi all'insolazione giornaliera anno per anno, e al valore giornaliero medio e minimo su un periodo di 4 anni (in $kwh/m^2/giorno$), osservati lungo un mese di durata 30 giorni in un caso didattico.

Osservando la Tabella 4.1 possiamo notare che il numero massimo di giorni consecutivi senza sole nel mese in considerazione è uguale a 3, semplicemente contando il numero di volte consecutive in cui osserviamo un'insolazione nulla. Nei casi reali, tuttavia, non avremo mai un'insolazione completamente nulla, ma ci saranno dei periodi, più o meno lunghi, durante i quali l'insolazione sarà al di sotto della media, e in quei periodi il sistema di backup dei pannelli fotovoltaici dovrà fornire l'energia che non arriva dall'insolazione.

Si rende quindi necessario un metodo per stimare il numero equivalente di giorni consecutivi senza sole anche quando l'insolazione disponibile non è nulla ma è minore della sua media.

Partendo dai dati in Tabella 4.1, possiamo innanzi tutto calcolare l'insolazione media mensile nel periodo in considerazione, che è uguale a $\frac{0.75*3+2*27}{30} = 1.875$. Quindi, $1.875 kwh/m^2/giorno$ è l'energia che vogliamo che l'impianto fornisca in maniera continua (che chiameremo nel seguito *energia attesa*).

La minima insolazione giornaliera invece è $0 kwh/m^2$ (cioè lo 0% dell'energia attesa). Questo vuol dire che la batteria o il sistema di backup deve poter fornire, nel giorno di insolazione nulla, il 100% dell'energia desiderata, cioè $1.875 kwh/m^2$. Stimiamo quindi che, su periodi consecutivi di un giorno, il numero equivalente di giorni senza sole è 1,

poiché la batteria deve poter fornire un quantitativo di energia pari all'energia media attesa in un giorno.

A questo punto, quindi, consideriamo periodi di due giorni consecutivi. Su due giorni, l'energia che l'impianto deve fornire è $1.875 * 2 = 3.75 kwh/m^2$. L'insolazione minima per due giorni consecutivi è ancora 0, quindi la batteria deve fornire, su due giorni, il 100% dell'energia attesa, cioè $3.75 kwh/m^2$, che è l'equivalente di due giorni di funzionamento dell'impianto con insolazione media. Quindi, il numero equivalente di giorni senza sole su periodi di due giorni è 2.

Lo stesso calcolo si può fare per periodi di tre giorni consecutivi, poiché l'insolazione minima è ancora 0, ottenendo quindi che il numero equivalente di giorni senza sole su periodi di tre giorni è 3.

Passiamo ora a periodi di quattro giorni consecutivi. L'energia che l'impianto deve fornire è $1.875 * 4 = 7.5 kwh/m^2$. L'insolazione minima disponibile su periodi consecutivi di quattro giorni è invece uguale a $2 kwh/m^2$ (corrispondente a un periodo di 4 giorni che contiene i 3 giorni di insolazione nulla), che corrisponde al 26.7% della quantità attesa. Quindi in quattro giorni la batteria deve fornire la percentuale di energia rimanente, cioè $5.5 kwh/m^2$, corrispondenti al 293.3% dell'energia corrispondente a un giorno. Quindi, il numero equivalente di giorni senza sole su periodi di quattro giorni è 2.93.

I risultati del calcolo del numero equivalente di giorni senza sole consecutivi su periodi di 1,2,3,4,20 e 30 giorni è riassunto in Tabella 4.2.

giorni consecutivi	ins. minima	ins. attesa	% mancante	giorni consecutivi senza sole
1	0	1.875	100%	1
2	0	3.75	100%	2
3	0	5.625	100%	3
4	2	7.5	73.3%	2.93
20	34	37.5	9.3%	1.86
30	54	56.25	4%	1.2

Tabella 4.2: Calcolo del numero equivalente di giorni consecutivi senza sole su finestre di 1,2,3,4,5,20 e 30 giorni in un caso limite

Osserviamo che, aumentando il periodo considerato da 3 a 4 giorni, la quantità stimata inizia a diminuire, poiché su periodi più lunghi i dispositivi si possono ricaricare a causa della presenza di giorni in cui l'insolazione incidente è maggiore dell'insolazione media (nel nostro esempio, nel periodo di 4 giorni, il giorno con insolazione uguale a $2 kwh/m^2$). Per stimare il numero equivalente di giorni consecutivi senza sole quindi, in linea teorica, è necessario effettuare il calcolo descritto su finestre temporali di lunghezza qualsiasi (anche non multipli interi di giorni), e poi calcolare il massimo dei valori ottenuti.

Nella pratica, tale calcolo sarebbe impossibile poiché, volendo considerare ogni possibile finestra temporale, dovremmo effettuare una quantità non numerabile di operazioni, per poi stimare il massimo. Quello che, invece, è stato fatto dalla NASA, è calcolare il numero equivalente di giorni senza sole su finestre di 1,3,7,14,21 e 30 giorni, e calcolare il massimo tra i valori così ottenuti. La grandezza risultante, che rappresenta una stima del numero equivalente di giorni consecutivi senza sole per un dato mese è quella che compone il dataset che analizzeremo. Nel caso didattico illustrato precedentemente, tale grandezza è esattamente uguale a 3 giorni.

Un'ultima osservazione va fatta riguardo i valori assunti da tale grandezza in corrispondenza dei siti della griglia ubicati al di sopra del circolo polare artico e al di sotto dell'antartico. In queste zone i dati relativi all'insolazione media mensile sono nulli per

almeno un mese all'anno. Infatti, anche se per latitudini poco superiori al circolo polare artico la notte polare dura solo pochi giorni, i dati dell'insolazione sono relativi all'insolazione incidente su di una superficie orizzontale posta sulla superficie terrestre. Nei mesi invernali al di sopra del circolo polare artico, i raggi del sole arrivano, per poche ore al giorno, con un angolo di incidenza molto piccolo, quindi l'effettiva insolazione disponibile se si posizionasse un pannello fotovoltaico orizzontalmente sulla superficie terrestre sarebbe nulla.

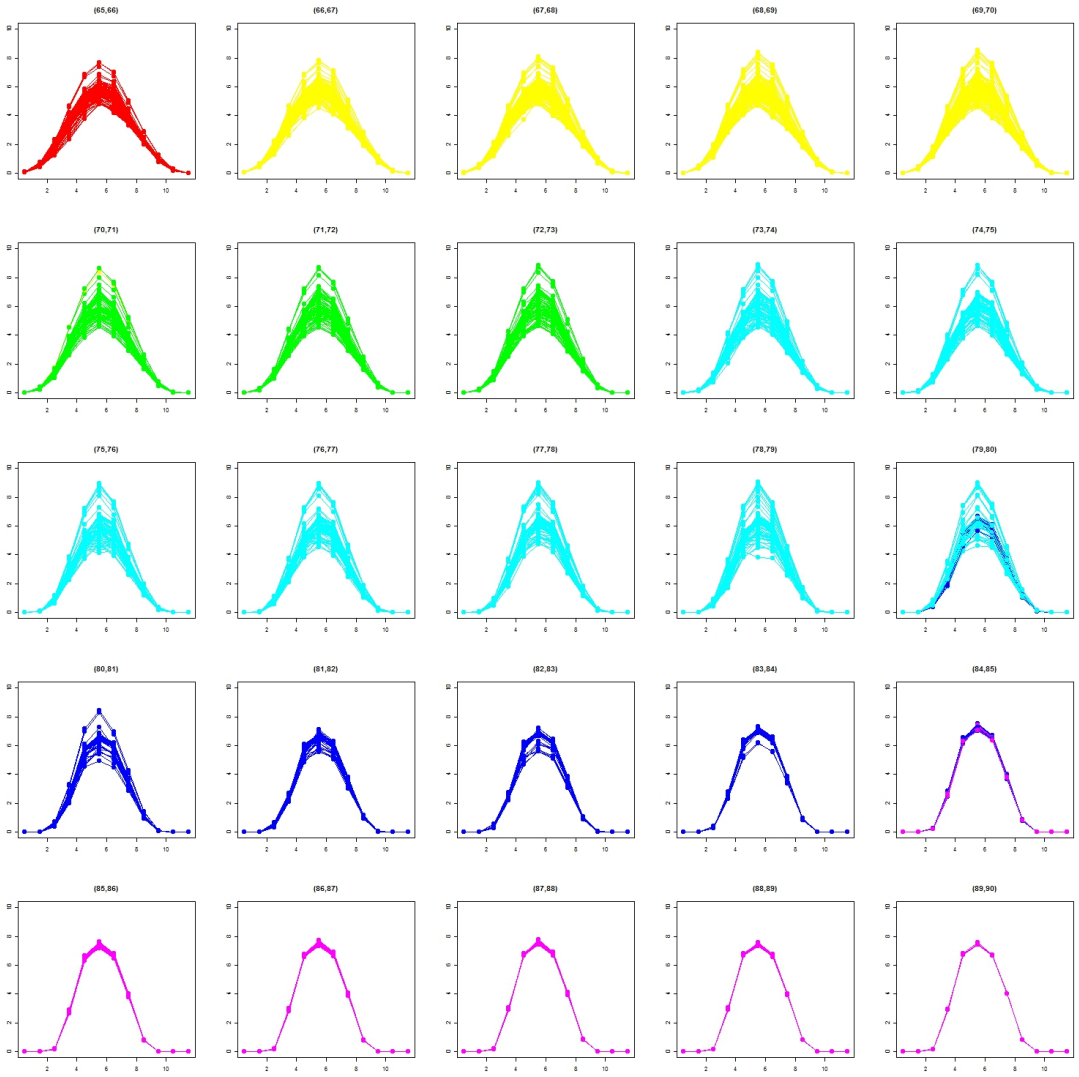


Figura 4.1: Dati relativi all'insolazione media mensile, al variare del mese lungo un anno, divisi per latitudine: da sinistra a destra e poi dall'alto in basso dati relativi a siti appartenenti ad intervalli di latitudine $A_{\lambda,65}$, $A_{\lambda,66}$, ...,$A_{\lambda,90}$. I colori rappresentano il numero di mesi a insolazione nulla: rosso=0 mesi, giallo=1 mese, verde=2 mesi, azzurro=3 mesi, blu=4 mesi, viola=5 mesi

In Figura 4.1 possiamo osservare l'insolazione media mensile per latitudini superiori a 65° (il circolo polare artico è nell'intervallo $(66, 67)°$); nella Figura, i colori delle curve indicano il numero di mesi in cui l'insolazione media è nulla: il colore rosso corrisponde a dati per i quali non c'è nessun mese a insolazione nulla, il giallo a dati con un mese a

insolazione nulla, verde per due mesi, blu per tre mesi, azzurro per quattro mesi e viola per cinque mesi.

La principale osservazione è che i mesi a insolazione nulla sono presenti esclusivamente a partire da 66° di latitudine. Come ci si poteva aspettare, poi, all'aumentare della latitudine, aumenta anche il numero di mesi a insolazione nulla. I dati osservati sono relativi al polo nord, ma analogamente si può vedere che al di sotto di $-66°$ c'è almeno un mese a insolazione nulla.

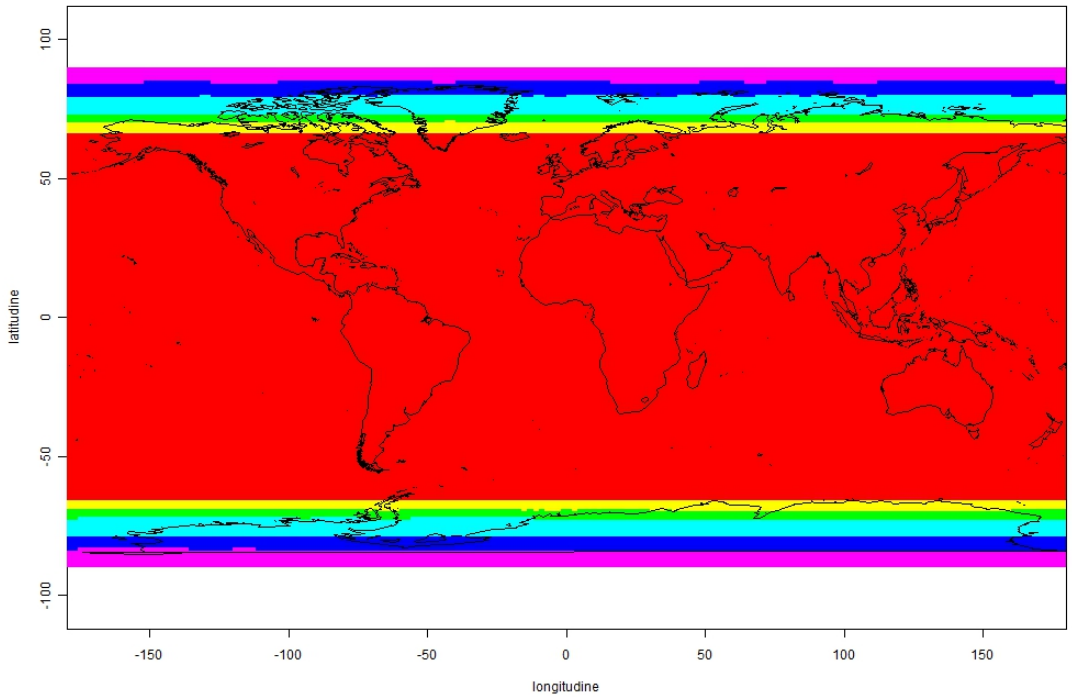


Figura 4.2: Aree del globo terrestre corrispondenti a dati in cui l'insolazione media è nulla per un certo numero di mesi: tutti i siti per i quali il numero di mesi ad insolazione media nulla è pari a 0 sono colorati in rosso, a 1 in giallo, a 2 in verde, a 3 in azzurro, a 4 in blu, a 5 viola.

In Figura 4.2 mostriamo infine la disposizione delle fasce corrispondenti ad un numero diverso di mesi con insolazione nulla sulla terra, dove i colori hanno lo stesso significato di prima. A dimostrazione di quanto detto precedentemente, i siti corrispondenti ad u numero di mesi a insolazione nulla compaiono esattamente sopra i 66° e sotto i $-66°$, in corrispondenza delle calotte polari.

Siamo interessati alla disposizione delle zone nelle quali l'insolazione è nulla per almeno un mese per una ragione molto semplice: ricordando come si calcola il numero equivalente di giorni consecutivi senza sole, notiamo che se l'insolazione media mensile è nulla, la quantità che siamo interessati ad analizzare non è definita, poiché per ogni finestra temporale l'insolazione minima disponibile e l'insolazione media sono entrambe nulle. Non possiamo quindi calcolare quanta percentuale dell'insolazione media rappresenta l'insolazione minima.

In base a quanto detto, quindi, la variabile che stiamo analizzando non è definita per alcuni mesi nelle calotte polari. Per la nostra analisi, quindi, dato che siamo interessati

a modellizzare l'andamento annuale di tale grandezza, scegliamo di estromettere i poli dall'analisi, considerando esclusivamente i dati compresi tra -66 e 66 gradi di latitudine.

4.2 Tecniche adottate per lo smoothing dei dati

4.2.1 Rappresentazione del dato

Come spiegato nella sezione precedente, i dati che vogliamo analizzare si riferiscono ad una stima del numero equivalente di giorni consecutivi senza sole. Abbiamo a disposizione tale stima per ogni mese dell'anno su un reticolo formato da meridiani e paralleli, nel quale un sito del reticolo corrisponde ad una regione di $1° \times 1°$, dal quale abbiamo escluso le aree appartenenti alle calotte polari (al di sopra di $66°$ e al di sotto di $-66°$).

Con la notazione dei capitoli precedenti, $N_1 = 132$ è il numero di righe del reticolo, mentre $N_2 = 360$ è il numero di colonne. Il numero totale di siti è quindi $N_1 \times N_2 = 47520$.

Per ogni sito $\mathbf{s}_i$ abbiamo a disposizione 12 valori $\tilde{y}_i(t_1), ..., \tilde{y}_i(t_{12})$, che immaginiamo essere realizzazioni puntuali della funzione aleatoria $\tilde{Y}_i(t), \quad t \in [0, 12]$.

C'è da fare una prima precisazione: l'unità di misura dei dati che abbiamo a disposizione è un numero di giorni al mese; tali dati vanno quindi normalizzati per ottenere grandezze tra loro confrontabili. Ad esempio, supponiamo di osservare $\tilde{y}_i(t_3) = \tilde{y}_i(t_4)$, cioè uno stesso valore per i mesi di marzo e aprile. Questo non vuol dire che il fenomeno che stiamo osservando non vari, tra marzo e aprile, poiché dobbiamo considerare il fatto che $\tilde{y}_i(t_3)$ può assumere valori tra 0 e 31, mentre $\tilde{y}_i(t_4)$ può valere al massimo 30. Prima di affrontare qualsiasi tipo di analisi, quindi, è necessario riscalare i dati relativi a mesi diversi, considerando, invece che il numero assoluto di giorni consecutivi senza sole $\tilde{Y}$, la proporzione di giorni consecutivi senza sole, così definita:

$$Y_i(t_j) = \frac{\tilde{Y}_i(t_j)}{g(t_j)} \quad j \in \{1, 2, ..., 12\}, \tag{4.1}$$

dove $g(t_j)$ rappresenta semplicemente il numero di giorni del mese j.

Per rappresentare il dataset composto dai dati riscalati che prendiamo in considerazione, mostriamo nelle Figure 4.3 e 4.4 delle mappe rappresentanti la proporzione di giorni senza sole consecutivi per ogni mese, nelle quali la proporzione di giorni relativa ad ogni sito è indicata dal colore del pixel ad esso associato. In particolare, utilizziamo una scala che va dal rosso, che indica siti in corrispondenza dei quali tale proporzione è quasi nulla, al blu, che indica siti in corrispondenza dei quali il parametro di riferimento è vicino al valore massimo osservato.

Dalle 12 mappe osserviamo che si intravede una struttura geografica secondo la quale si dispongono i dati. Tuttavia, l'alta variabilità del dataset rende necessario l'utilizzo di tecniche di classificazione che tengano conto anche delle informazioni spaziali al fine di ottenere un'identificazione nitida di regioni omogenee.

A questo punto, prima di procedere alla classificazione dei dati con le tecniche presentate, è necessario utilizzare metodi di smoothing per passare dalle realizzazioni puntuali che abbiamo a disposizione, a dati funzionali. Il dato funzionale associato al sito $\mathbf{s}_i$ deve rappresentare l'evoluzione nel tempo della proporzione di giorni senza sole. Supponiamo, quindi, che i dati puntuali a nostra disposizione, siano una stima di tale proporzione centrata su un dato mese, che siano cioè realizzazioni della quantità $Y_i(t_j)$ nei punti centrali di ogni mese: $t_1 = 0.5, t_2 = 1.5, ..., t_{12} = 11.5$.

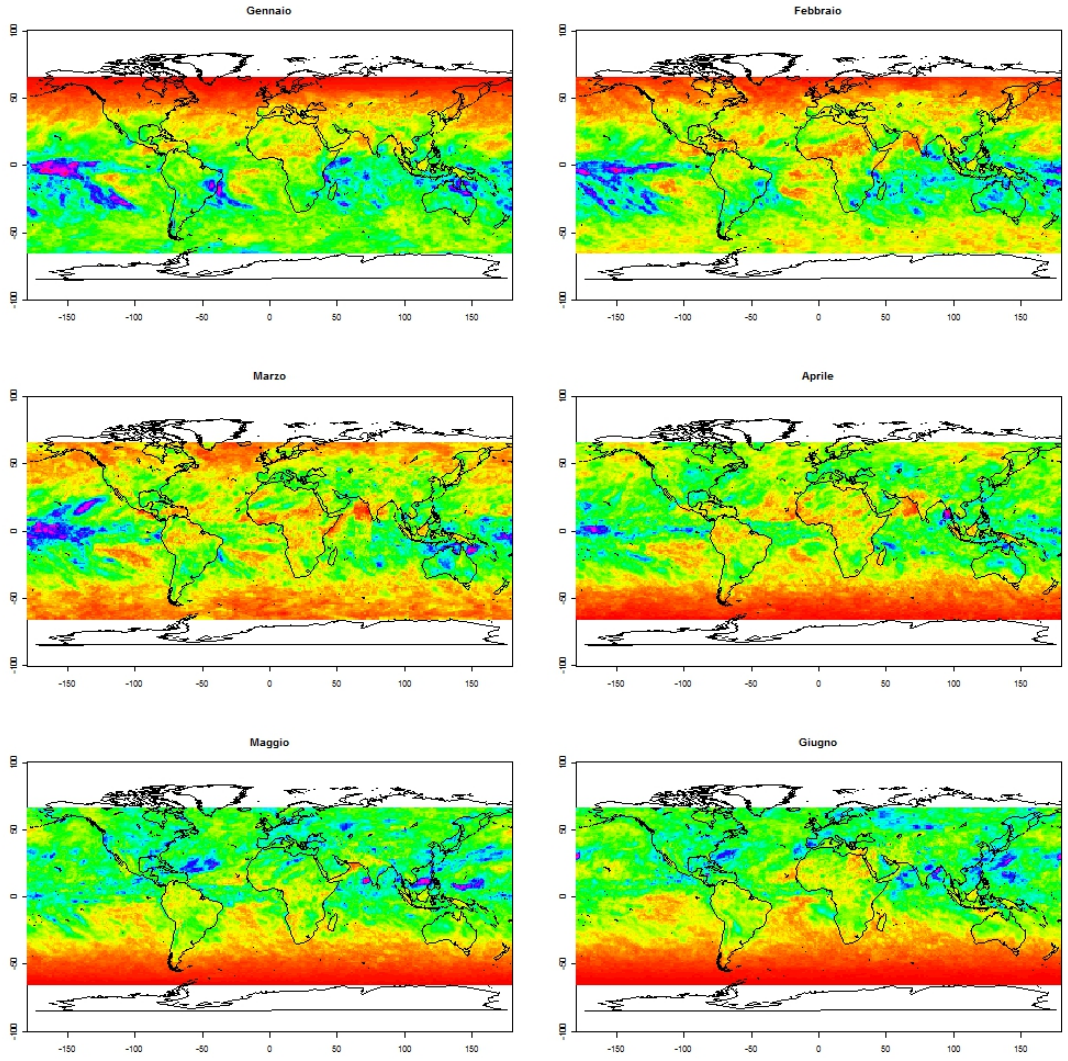


Figura 4.3: Mappe rappresentanti la proporzione mensile di giorni consecutivi senza sole osservati negli ultimi 22 anni per ogni sito, nei mesi da gennaio a giugno; in rosso vengono rappresentati valori vicini allo zero, in blu valori vicini al valore massimo osservato

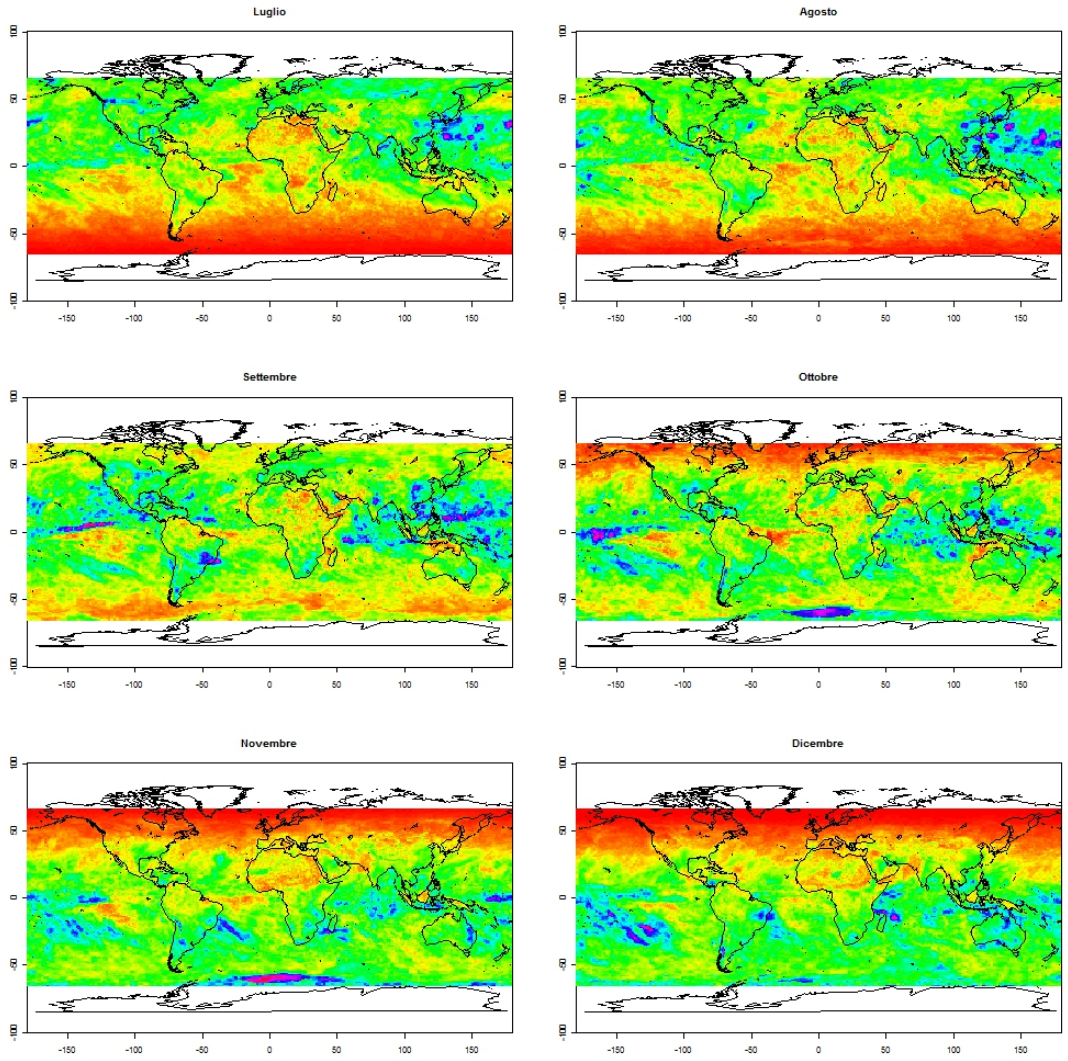


Figura 4.4: Mappe rappresentanti la proporzione mensile di giorni consecutivi senza sole osservati negli ultimi 22 anni per ogni sito, nei mesi da luglio a dicembre; in rosso vengono rappresentati valori vicini allo zero, in blu valori vicini al valore massimo osservato

In base alle informazioni sulla variabile considerata, per ogni sito $\mathbf{s}_i$, il dato funzionale y_i deve soddisfare due proprietà fondamentali:

1. periodicità: la funzione Y_i rappresenta un andamento annuale, quindi è naturale richiedere che sia periodica di periodo $T = 12$;

2. $Y_i(t) \in [0, 1] \quad \forall t \in [0, 12]$. La quantità $Y_i(t)$ è una proporzione, deve quindi essere compresa tra 0 e 1.

La proprietà 1. (periodicità) suggerisce l'idea di rappresentare i dati, come già era stato fatto per le simulazioni presentate nel Capitolo 3, come sviluppi, arrestati ad un determinato ordine, di serie di Fourier. I vantaggi dell'utilizzo di tale tecnica sono vari: innanzi tutto permette di effettuare la classificazione direttamente sui vettori di coefficienti della base, invece che sui dati funzionali stessi, senza dover effettuare tecniche di riduzione dimensionale del dato. L'utilizzo di tale rappresentazione permette inoltre di analizzare separatamente la possibile tendenza annuale dei dati (considerando solo le funzioni di periodo 12), l'andamento semestrale (considerando le funzioni di periodo 6) e così via. L'utilizzo di una base fissata a priori permette infine di non cambiare lo spazio di rappresentazione dei dati tra un'iterazione e l'altra del metodo presentato nel Capitolo 2. Infatti, la base composta dalle componenti principali dei dati di sintesi risultanti dalla tassellazione di Voronoi, è diversa ad ogni iterazione. Questa osservazione può costituire un problema, soprattutto nel caso in cui le funzioni di base trovate siano tra loro molto diverse.

Sfortunatamente, però, l'utilizzo di una base fissa, come quella di Fourier, non garantisce che sia rispettata la proprietà 2., cioè che i dati funzionali siano compresi tra 0 e 1. Tale proprietà non può essere garantita, e nel nostro caso utilizzare la base di Fourier per rappresentare i dati porta alla sua violazione.

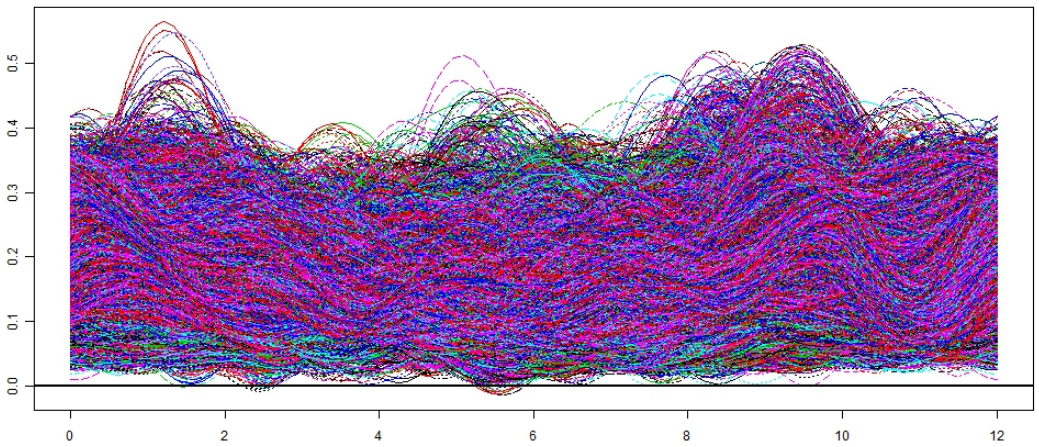


Figura 4.5: Dati funzionali ottenuti tramite smoothing con base di Fourier di dimensione $p = 11$

Osserviamo infatti in Figura 4.5 il risultato ottenuto utilizzando per la rappresentazione dei dati funzionali le prime 11 funzioni della base di Fourier. Sebbene la maggior parte dei dati rispetti la proprietà 2., vediamo che alcuni di essi hanno delle oscillazioni al di sotto della linea delle ascisse, assumendo valori negativi per alcuni valori di t.

Scegliamo quindi di utilizzare un approccio più generale, effettuando lo smoothing con l'utilizzo di un kernel gaussiano, che descriviamo in dettaglio nel Paragrafo 4.2.2 (si veda anche [Härdle, 1991]), che, come vedremo, garantisce che la proprietà 2. venga rispettata, per poi effettuare la classificazione come spiegato nel Capitolo 2, per mezzo di FPCA e classificazione sugli scores.

4.2.2 Smoothing per mezzo di un kernel gaussiano

Immaginiamo, in generale, di avere a disposizione delle osservazioni della funzione aleatoria $Y(t) : [t_{min}, t_{max}] \to \mathbb{R}$ nei punti $t_1, ..., t_J$, e di voler ricostruire il dato funzionale, cioè calcolare il valore della funzione in ogni punto t del dominio (tecnica detta smoothing del dato funzionale).

Effettuare smoothing per mezzo di un kernel significa calcolare il valore del dato funzionale $y_i(t)$ in ogni punto $t \in [t_{min}, t_{max}]$ come media pesata dei valori osservati nei punti $t_j, \quad j = 1, ..., J$. I pesi associati agli $y(t_j)$ dipendono dalla distanza di t_j dal punto t nel quale si vuole calcolare il valore della funzione, in modo da pesare di più i dati più vicini al punto t, e di meno quelli più lontani.

In particolare, la stima di $y(t)$ è data da una combinazione lineare delle osservazioni:

$$\hat{y}(t) = \sum_{j=1}^{J} S_j(t) y(t_j), \tag{4.2}$$

per un'opportuna funzione di pesi $S_j(t)$. La forma più utilizzata per il calcolo del valore $\hat{y}(t)$ è la stima di Nadaraya-Watson, che utilizza:

$$S_j(t) = \frac{K[(t_j - t)/h]}{\sum_{r=1}^{J} K[(t_r - t)/h]}, \tag{4.3}$$

dove $K(u)$ è la *funzione di kernel*, non negativa, simmetrica e centrata in 0, mentre il parametro h è detto *bandwidth* e regola la forma della funzione risultante.

La formula 4.3 per la stima di $y(t)$ tramite 4.2, pur rimanendo una forma molto generale (può essere utilizzata con funzioni di kernel con proprietà molto diverse), ha la proprietà che, per t fissato, i pesi $S_j(t)$ associati alle osservazioni $y(t_j)$ sommano a uno:

$$\sum_{j=1}^{J} S_j(t) = 1 \quad \forall t \in [t_{min}, t_{max}].$$

Questa proprietà garantisce che il range entro cui varia la funzione stimata $\hat{y}(t)$ sia compreso tra la massima e la minima osservazione:

$$\min_j y(t_j) \leq \hat{y}(t) \leq \max_j y(t_j) \quad \forall t \in [t_{min}, t_{max}]. \tag{4.4}$$

L'equazione 4.4 garantisce, banalmente, che se le osservazioni sono comprese nell'intervallo $[0, 1]$, come nel caso preso in esame, anche la funzione stimata assuma valori in $[0, 1]$, rispettando la proprietà 2. che abbiamo posto come restrizione per il dato funzionale $y(t)$.

Nel nostro caso, poi, scegliamo di effettuare lo smoothing utilizzando un kernel gaussiano, cioè definendo:

$$K(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}}. \tag{4.5}$$

Questa scelta permette di ottenere funzioni $\hat{y}(t)$ continue e derivabili infinite volte nel dominio (t_{min}, t_{max}), poiché ottenute come misture di distribuzioni normali centrate nei dati osservati, con varianza uguale al parametro di bandwidth h. La formula esplicita per il calcolo di $\hat{y}(t)$ è dunque:

$$\hat{y}(t) = \frac{\sum_{j=1}^{J} e^{-(t_j - t)^2/2h^2} y(t_j)}{\sum_{r=1}^{J} e^{-(t_j - t)^2/2h^2}}. \tag{4.6}$$

Dall'espressione 4.6 capiamo come agisca sullo smoothing il bandwidth h, che rappresenta la varianza della distribuzione normale associata ai dati. Per valori elevati del bandwidth, osservazioni relative a diversi istanti di tempo t_j hanno pesi molto simili, poiché la distanza $t - t_j$ è divisa per un fattore molto grande. Il dato funzionale risultante sarà quindi piatto, senza picchi considerevoli, e si discosterà molto dalle osservazioni nei punti t_j. Al contrario, per valori molto bassi del parametro h la distanza tra i punti gioca un ruolo maggiore, e il peso associato a osservazioni molto distanti dal punto t sarà molto piccolo. La funzione $\hat{y}(t)$ risulterà essere quasi interpolante per le osservazioni, con variazioni nette tra un valore osservato e l'altro.

La scelta del bandwidth dipende dalla variabilità dei dati in esame. Nel nostro caso, effettueremo lo smoothing per diversi valori di h per poi scegliere il valore che permetta una migliore rappresentazione dei dati.

Rimane da imporre la periodicità del dato funzionale, che non è garantita dalla stima proposta. Bisogna inoltre tener conto del fatto che le tecniche di smoothing che utilizzano una media pesata delle osservazioni a disposizione presentano problemi sui bordi del dominio. Infatti, mentre per stimare il valore di $y(t)$ in un punto centrale si calcola una combinazione lineare di valori osservati in un intervallo simmetrico centrato sul punto t, ai bordi del dominio questo non è possibile poiché si ha a disposizione esclusivamente un intervallo destro o sinistro di t.

Il metodo che utilizziamo per rendere la funzione $\hat{y}(t)$ approssimativamente periodica, e contemporaneamente ovviare al problema brevemente descritto, è molto semplice: scegliamo di orlare il dominio di riferimento, calcolando la stima $\hat{y}_o(t)$ partendo da un vettore di osservazioni di dimensione $J = 36$ che consiste nella ripetizione per tre volte dell'intero vettore di osservazioni. Stimiamo quindi la funzione $y(t)$ in un dominio pari a tre volte il suo periodo: il dominio orlato sul quale valutiamo la funzione diventa quindi $[t_{min}, t_{max}] = [0, 36]$. Una volta ottenuta in questo modo la stima $\hat{y}_o(t), \quad t \in [0, 36]$, andiamo a considerare come dato funzionale esclusivamente la parte della funzione stimata nel periodo centrale, traslandola nell'origine, cioè poniamo:

$$\hat{y}(t) = \hat{y}_o(t + 12) \quad \forall t \in [0, 12].$$

In questo modo forziamo il dato funzionale $\hat{y}(t)$ ad essere una funzione approssimativamente periodica, ed eliminiamo eventuali problemi di instabilità ai bordi del dominio della stima ottenuta utilizzando una tecnica di smoothing con kernel.

4.2.3 Risultati ottenuti e scelta del bandwidth

Nei paragrafi precedenti abbiamo fornito una motivazione per la scelta del metodo utilizzato per lo smoothing dei dati e della funzione di kernel utilizzata (ricordiamo che vogliamo effettuare uno smoothing con kernel gaussiano).

Osserviamo che dobbiamo effettuare lo smoothing sulle osservazion $y_i(t_j)$ relative ad ogni sito $\mathbf{s_i}$. Per una questione di coerenza, scegliamo di utilizzare la stessa tecnica

di smoothing e lo stesso bandwidth per ogni sito. L'unica cosa che rimane da fissare, è quindi il parametro di bandwidth h.

Per scegliere il bandwidth, effettuiamo lo smoothing di tutti i dati a nostra disposizione per diversi valori del parametro di bandwidth, per poi scegliere quello che permette di ottenere risultati migliori in termini di rappresentazione del dato. Idealmente, possiamo supporre che una buona rappresentazione del dato debba essere simile a quanto avevamo ottenuto utilizzando la base di Fourier (Figura 4.5), eliminando i problemi di negatività che avevamo riscontrato nell'utilizzo di tale metodo.

In Figura 4.6 mostriamo il risultato dello smoothing per diversi valori di h, da 0.5 a 3, effettuato su di un campione casuale di 18 siti. I grafici mostrano, oltre al risultato dello smoothing, anche i valori delle osservazioni $y(t_j)$. Notiamo innanzi tutto che i dati funzionali ottenuti seguono il comportamento che avevamo previsto: per valori piccoli del bandwidth, le funzioni che otteniamo seguono la successione delle osservazioni $y(t_j)$, e si osservano variazioni brusche da un valore osservato all'altro. Al limite, per $h \to 0$ otteniamo funzioni costanti a tratti. Al contrario, per valori elevati di h, sempre più osservazioni entrano nel calcolo del valore $y(t)$. Le funzioni risultanti hanno quindi un andamento più liscio, per diventare al limite per $h \to \infty$ costanti e uguali alla media delle osservazioni.

Quello che vorremmo, per una buona rappresentazione dei dati come funzioni, è un valore di h intermedio tra le due situazioni limite: non è necessario che il dato funzionale sia interpolante per le osservazioni, vogliamo però che segua l'andamento dei dati osservati in maniera liscia, senza variazioni troppo brusche. Un'altra informazione che possiamo utilizzare per la scelta di h è un confronto con i dati ottenuti con lo sviluppo di Fourier, in Figura 4.5.

Osserviamo che, confrontando i vari grafici, il valore del bandwidth che sembra fornire risultati più simili a quelli ottenuti con lo sviluppo di Fourier è $h = 1.5$. Osserviamo che, per tale parametro, i dati funzionali non interpolano le osservazioni, ma ne seguono l'andamento in maniera liscia. Ricordiamo infatti che stiamo assumendo che le osservazioni mensili del parametro che stiamo modellizzando, non sono certe, ma costituiscono realizzazioni puntuali di una funzione aleatoria: sono quindi affette da rumore. Un dato funzionale interpolante delle osservazioni, quindi, non è significativo per questo tipo di modello. Una buona tecnica di smoothing dei dati, ha come obiettivo proprio quello di ricostruire il dato funzionale partendo dalle osservazioni puntuali in modo da eliminare, o ridurre, la componente di rumore delle osservazioni $y(t_j)$. A questo scopo, per la stima di $y(t)$ in ogni punto del dominio si utilizza il di osservazioni $\{y(t_j), j = 1, ..., J\}$. In particolare, $\hat{y}(t_j) \neq y(t_j)$, e la funzione stimata non è interpolante, poiché anche in corrispondenza dei punti osservati t_j la stima tiene conto dell'intero vettore di osservazioni.

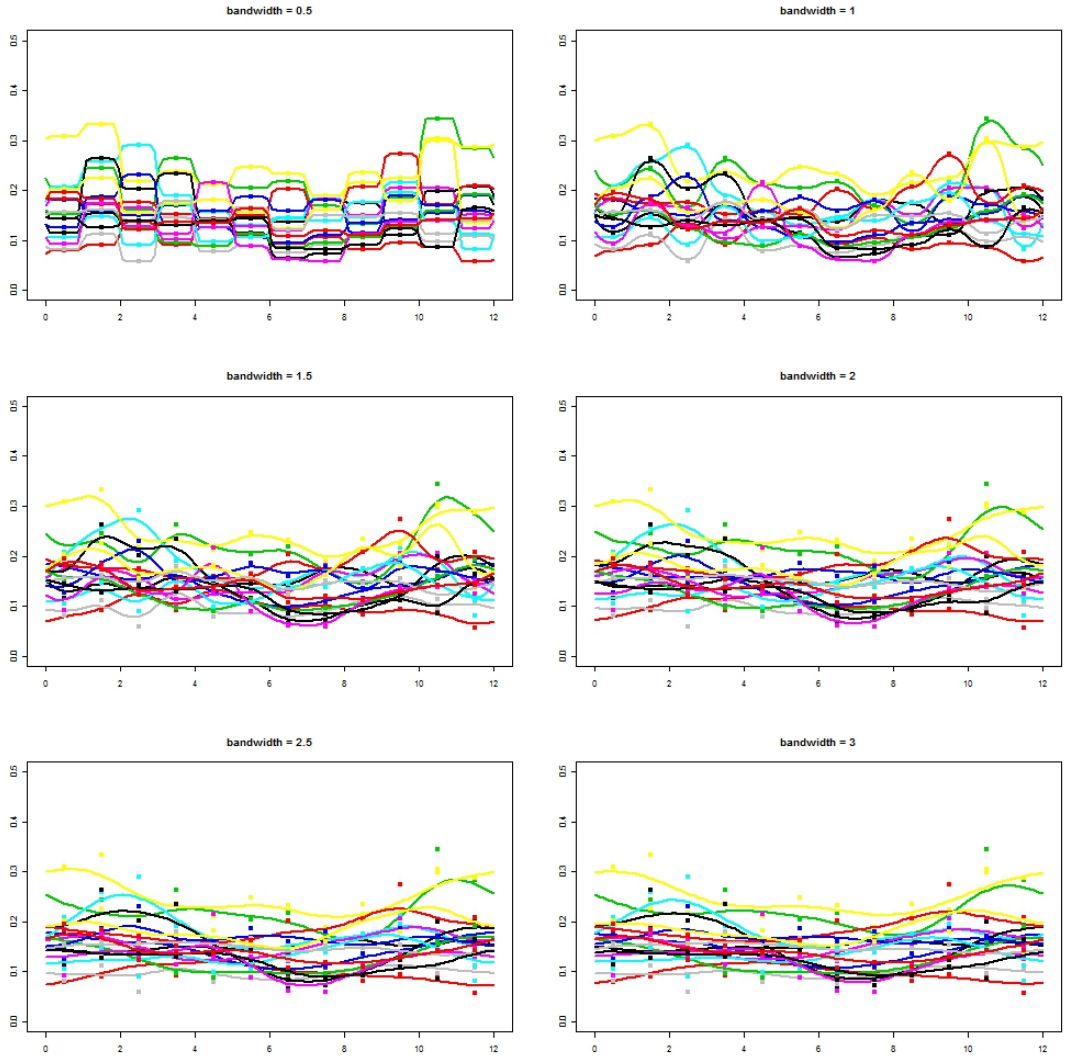


Figura 4.6: Dati funzionali ottenuti tramite smoothing con kernel gaussiano per 18 siti estratti casualmente dal dataset, da sinistra a destra e poi dall'alto in basso con bandwith $h = 0.5, 1, ..., 3$

4.3 Spatial clustering dei dati con k–medie

Il dataset che abbiamo creato con le tecniche di smoothing presentate nella Sezione 4.2 si situa esattamente nel contesto generale proposto: abbiamo un reticolo regolare di $N_1 \times N_2$ siti, e per ciascun sito $\mathbf{s}_i \in \mathcal{S}$ abbiamo a disposizione un dato funzionale y_i.

L'obiettivo dell'analisi è la classificazione non supervisionata di tale dataset. Vogliamo cioè associare ad ogni sito un'etichetta $l \in \mathcal{L}$ (dove $\mathcal{L}$ è un insieme discreto di cardinalità k) in modo che dati associati alla stessa etichetta abbiano caratteristiche comuni.

L'analisi che ci proponiamo di effettuare rappresenta un buon caso studio per l'utilizzo delle tecniche proposte, in quanto le dimensioni del dataset che consideriamo sono abbastanza ridotte: è possibile, nel nostro caso, sia classificare i dati iterando la riduzione del dataset originale e classificazione dei dati di sintesi (spatial clustering) che effettuare una classificazione non spaziale tramite k–medie (simple clustering), al fine di poter effettuare un confronto tra i risultati ottenuti con i due metodi, come già mostrato per gli studi di simulazione del Capitolo 3.

I dati funzionali che abbiamo a disposizione rappresentano l'evoluzione nel tempo di una determinata grandezza, legata a particolari caratteristiche climatiche, e cioè la proporzione di giorni senza sole consecutivi. Il risultato della classificazione di questi dati, quindi, metterà in evidenza le zone del mondo in cui l'evoluzione di tale grandezza ha caratteristiche comuni.

É naturale supporre, vista la natura dei dati, che ci sia una dipendenza spaziale tra i dati stessi, modellizzabile attraverso uno dei modelli descritti nel Capitolo 1; quindi la tecnica di classificazione utilizzata dovrà tenere conto di tale dipendenza, come proposto nel Capitolo 2.

Osserviamo che la tecnica di classificazione proposta nel Capitolo 2 e successivamente analizzata tramite studi di simulazione, si basa su un'idea generale, che deve essere adattata al contesto dei dati in analisi. Prima di tutto, quindi, è necessario declinare l'idea proposta, nel caso che stiamo trattando.

Ricordiamo brevemente che l'algoritmo generale che abbiamo proposto consiste in tre fasi principali:

- tassellazione di Voronoi e calcolo dei dati di sintesi dei tasselli;
- riduzione dimensionale del dataset composto dai dati di sintesi dei tasselli;
- classificazione dei dati di sintesi dei tasselli.

Per effettuare la classificazione dei dati, quindi, è necessario definire la distanza tra i siti che verrà utilizzata per trovare i diagrammi di Voronoi, una tecnica di riduzione dimensionale ed una di classificazione multivariata da utilizzare.

Oltre a queste scelte, riguardanti la struttura dell'algoritmo, è necessario fissare alcuni parametri. In particolare, bisogna scegliere il numero di tasselli del diagramma di Voronoi n, il numero di iterazioni dell'algoritmo M ed il numero di cluster k. Per quanto riguarda la scelta del numero di cluster, utilizzeremo il metodo proposto analizzando l'entropia delle classificazioni ottenute (Paragrafo 4.3.4). Per quanto invece riguarda gli altri due parametri, scegliamo di effettuare le analisi con $n = 300$ e $M = 100$.

Nei prossimi paragrafi, considereremo invece le scelte fatte per quanto riguarda la struttura dell'algoritmo proposto, cioè la scelta della distanza, dei metodi di riduzione dimensionale e dei metodi di classificazione, analizzando passo per passo i risultati ottenuti grazie a tali scelte.

4.3.1 Distanza tra i siti e tassellazione di Voronoi

Prima di procedere nel definire il tipo di metodo utilizzato per simulare, nel caso preso in esame, una tallessazione di Voronoi, è necessario analizzare accuratamente la geometria del reticolo nel quale il dataset è suddiviso. Infatti, sebbene il dataset si presenti esattamente come i dataset sintetici che abbiamo creato nel Capitolo 3, cioè osservazioni funzionali in corrispondenza di un reticolo regolare, nel caso preso in esame il reticolo è creato su di una superficie sferica, invece che sul piano. Utilizzare, quindi, la distanza euclidea tra i siti per definire una tassellazione di Voronoi sarebbe un errore, poiché tale distanza non rappresenta la reale distanza esistente tra i diversi punti posti sulla superficie terrestre.

Un'altra differenza fondamentale tra un dataset relativo a un reticolo sul piano e il dataset considerato è che, nel primo caso ogni quadrato o rettangolo in cui si suddivide il reticolo occupa un'uguale superficie sul piano; nel secondo caso, invece, i rettangoli che si formano non hanno, come vedremo, una superficie costante. Ogni dato funzionale, quindi, contiene una quantità di informazione relativa a un'area geografica di superficie diversa, e sarà necessario tenere conto di questa osservazione nel seguito dell'analisi.

All'inizio di questo paragrafo, quindi, analizziamo brevemente la geometria del reticolo considerato, mostrando, in particolare, che le superfici dei rettangoli $A_{\lambda,\theta}$ variano in funzione della latitudine θ. Successivamente, proporremo una tecnica di tassellazione della superficie terrestre che tenga conto di quanto detto.

Geometria del reticolo nel quale è suddiviso il dataset

L'obiettivo che ci poniamo è mostrare che ogni sito del reticolo del tipo $A_{\lambda,\theta}$ si riferisce ad aree di dimensione diversa sulla superficie terrestre. In particolare, vogliamo calcolare la superficie di $A_{\lambda,\theta}$ (che indicheremo con $S(A_{\lambda,\theta})$) per ogni valore intero di λ e θ, dove per effettuare il calcolo, supponiamo, per semplicità, che la terra sia una sfera di raggio R. In Figura 4.7 mostriamo degli esempi di aree delle quali vogliamo calcolare la superficie, ovvero porzioni della superficie sferica formate dalle intersezioni di meridiani e paralleli. In generale, il problema che ci proponiamo di risolvere è un problema di calcolo di superfici su di una sfera.

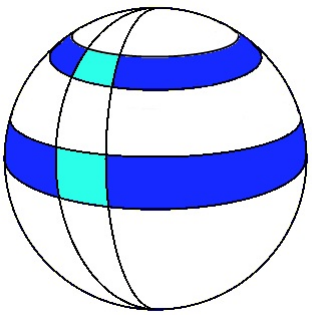


Figura 4.7: Esempi di aree sulla superficie terrestre formate dal'intersezione tra meridiani e paralleli

Ricordiamo, innanzi tutto, in significato delle coordinate geografiche (λ, θ), cioè latitudine e longitudine di un punto P sulla superficie terrestre. La latitudine θ è la misura in gradi dell'angolo che il segmento formato congiungendo il punto P con il centro della terra O forma con il piano equatoriale. La longitudine λ è invece la misura

in gradi dell'angolo che la proiezione di tale segmento sul piano equatoriale forma con il meridiano di Greenwhich.

Per il calcolo delle superfici su di una sfera, utilizziamo il seguente:

Teorema 4.1. *La proiezione delle coordinate geografiche sul piano che si ottiene nel modo seguente:*

$$\begin{cases} x &= \lambda \\ y &= \sin\theta \end{cases} \tag{4.7}$$

conserva le aree.

Nel sistema di equazioni che definisce la proiezione su piano 4.7 (che chiameremo nel seguito *proiezione cilindrica ad aree uguali*), (λ, θ) sono le due coordinate geografiche, mentre x e y rappresentano le coordinate nel nuovo sistema di riferimento piano, dove l'asse delle ascisse rappresenta l'equatore mentre l'asse delle ordinate il meridiano di Greenwich.

Il significato geometrico della proiezione 4.7 è il seguente: per proiettare sul piano i punti della sfera, innanzi tutto li proiettiamo sulla superficie laterale di un cilindro C di raggio R ed altezza $2R$ tangente alla superficie terrestre in corrispondenza dell'equatore, come in Figura 4.8. Successivamente, basta svolgere la superficie cilindrica, per ottenere la mappa piana. Il risultato è una mappa nella quale i meridiani ono tra loro equidistanti, mentre le distanze tra i paralleli diminuiscono man mano che ci si avvicina ai poli. Il Teorema 4.1 è importante per la ragione seguente: invece di calcolare la superficie di $A_{\lambda,\theta}$ sulla sfera, è possibile calcolare la misura di tale superficie proiettata sul cilindro, e il risultato numerico resta lo stesso.

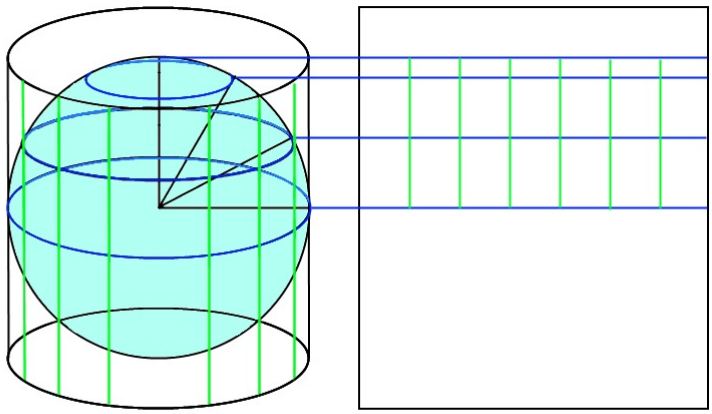


Figura 4.8: Schema illustrativo della costruzione della proiezione cilindrica ad aree uguali

Supponiamo, prima di tutto, di suddividere la superficie terrestre in strisce, utilizzando i paralleli a latitudini intere. Una volta calcolata la superficie compresa tra due paralleli successivi, basterà suddividerla anche lungo i meridiani, cioè dividere la superficie ottenuta per 180 per ottenere la superficie $S(A_{\lambda,\theta})$ cercata. In particolare, suddividiamo la terra in *segmenti sferici* così definiti: un segmento sferico è il solido che

si ottiene tagliando una sfera con due piani paralleli (si veda la Figura 4.9). Nel caso preso in esame, consideriamo due piani paralleli all'equatore. Se i due piani vengono

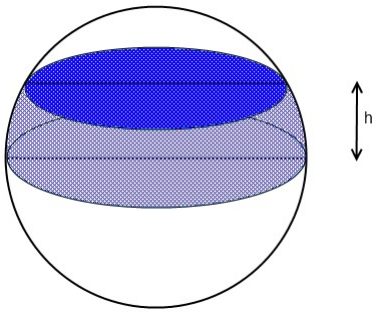


Figura 4.9: Segmento sferico ottenuto tagliando una sfera di raggio R con due piani paralleli all'equatore di distanza h

presi in modo da intersecare la superficie della sfera in corrispondenza di due paralleli di latitudine θ_1 e $\theta_2 = \theta_1 + 1$, la superficie laterale del segmento sferico (che chiamiamo A_{θ_1}) rappresenta esattamente la porzione di superficie terrestre compresa tra due paralleli successivi.

Per calcolare la superficie del segmento sferico A_{θ_1}, utilizziamo il Teorema 4.1. Infatti, la superficie che vogliamo calcolare equivale alla superficie laterale del cilindro C compresa tra i due piani che identificano i paralleli (Figura 4.10). Se chiamiamo h la

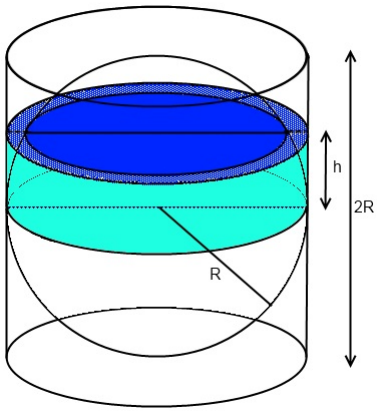


Figura 4.10: Corrispondenza tra la superficie laterale del segmento sferico ottenuto tagliando una sfera di raggio R con due piani paralleli all'equatore di distanza h e la superficie laterale del cilindro C tagliato dagli stessi piani

distanza tra i due piani, tale superficie è quindi uguale a:

$$S(A_{\lambda_1}) = 2\pi Rh. \tag{4.8}$$

Resta da trovare l'espressione per la distanza tra i due piani h. Dalla definizione che abbiamo dato delle coordinate geografiche, abbiamo che la distanza dal piano equatoriale di un punto con latitudine θ è data da $d = R\sin\theta$. Quindi, la distanza tra due piani paralleli al piano equatoriale che intersecano la superficie terrestre in corrispondenza di due paralleli successivi è data da:

$$h = R|\sin\theta_1 - \sin\theta_2| = R|\sin\theta - \sin(\theta + 1)|. \tag{4.9}$$

In conclusione, la superficie $S(A_{\lambda,\theta})$ cercata è data da:

$$S(A_{\lambda,\theta}) = \frac{2\pi R^2 |\sin\theta - \sin(\theta+1)|}{180}. \tag{4.10}$$

Dall'equazione 4.10 notiamo che, come potevamo aspettarci, l'espressione della superficie cercata dipende dalla latitudine θ ma non dalla longitudine λ. Inoltre, tale superficie, è maggiore nelle zone equatoriali, cioè per valori di θ vicini allo zero, ed è minore in corrispondenza dei poli.

Tassellazione di Voronoi sulla sfera

Proponiamo ora un metodo per effettuare una tassellazione di Voronoi sulla sfera. Per generare una tassellazione abbiamo bisogno di definire il metodo con il quale estraiamo casualmente i centri e la distanza da utilizzare per la tassellazione.

Per quanto riguarda la scelta della distanza tra i siti, osserviamo che nel dataset che abbiamo a disposizione i siti $\mathbf{s}_i$ sono aree geografiche di un grado di latitudine per un grado di longitudine (che nel seguito rappresenteremo utilizzando esclusivamente il punto centrale di tale area, ovvero indicheremo l'area $A_{\lambda,\theta} = [\lambda, \lambda+1] \times [\theta, \theta+1]$ con le coordinate $\mathbf{s}_{\lambda,\theta} = (\lambda + 0.5, \theta + 0.5)$) su tutta la terra, escluse le calotte polari.

La distanza che utilizzeremo è quindi la distanza geodetica, che rappresenta il minimo percorso sulla superficie terrestre (che supponiamo essere una sfera di raggio $R = 6371km$) che congiunge due punti. La distanza geodetica tra due siti di coordinate (latitudine, longitudine) $\mathbf{s}_1 = (\lambda_1, \theta_1)$ e $\mathbf{s}_2 = (\lambda_2, \theta_2)$ si ottiene come:

$$d(\mathbf{s}_1, \mathbf{s}_2) = R \arccos[\sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2 \cos(\lambda_1 - \lambda_2)]. \tag{4.11}$$

Per trovare una tassellazione di Voronoi, quindi, scegliamo di generare uniformemente n siti (centri) sulla superficie terrestre, e associamo ogni altro sito al centro più vicino. Procediamo poi ad eliminare dalla mappa così ottenuta i tasselli (o le parti dei tasselli) che si trovano nelle zone polari, poiché in quelle zone non abbiamo a disposizione dati funzionali. Un possibile risultato di tale procedura è mostrato in Figura 4.11.

Osserviamo che nella rappresentazione in Figura 4.11, i tasselli vicini alle zone polari sembrano più grandi di quelli nella zona equatoriale. Sappiamo infatti che la superficie di una sfera non può essere proiettata su di un piano in modo da conservare le distanze tra i punti, ovvero è impossibile rappresentare su di un piano la superficie terrestre senza introdurre distorsioni. Per questo motivo, sono state introdotte in letteratura (si veda [Banerjee et al., 2004]) diverse proiezioni utilizzate comunemente per rappresentare sul piano la superficie terrestre: alcune di esse, come la proiezione 4.7 introdotta precedentemente conservano le aree, altre gli angoli, altre ancora non hanno particolari proprietà geometriche, ma vengono utilizzate perché più semplici da definire. La mappa in Figura 4.11 è in particolare il risultato di una *proiezione equirettangolare* della superficie terrestre. Tale proiezione, consiste semplicemente nel rappresentare in ascissa la longitudine e in ordinata la latitudine di ogni punto sulla superficie terrestre:

$$\begin{cases} x &= \lambda \\ y &= \theta \end{cases} \tag{4.12}$$

La proiezione utilizzata non ha proprietà geometriche rilevanti. Utilizzando tale proiezione, infatti, né le aree né gli angoli o le forme vengono conservati. Tuttavia, si

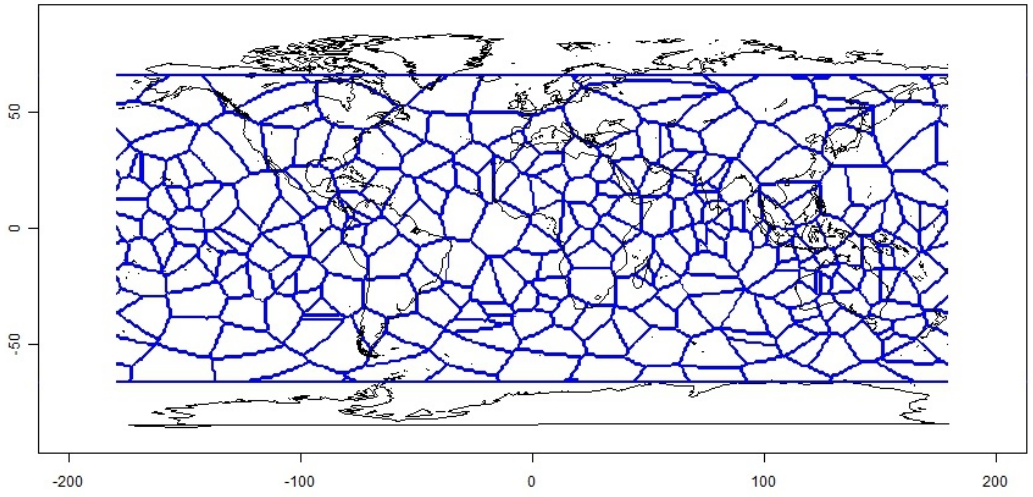


Figura 4.11: Una possibile tassellazione di Voronoi basata sulla distanza geodetica fra siti, ottenuta a partire da $n = 300$ centri generati casualmente in maniera uniforme sulla superficie terrestre e dopo aver eliminato le zone polari

tratta di una proiezione molto semplice da utilizzare. Osserviamo che, inoltre, proiettando in questo modo la superficie terrestre, ogni sito della griglia formata da meridiani e paralleli utilizzata dalla NASA per la rilevazione dei dati (corrispondente all'area $A_{\lambda\theta}$), corrisponde ad un'area della stessa dimensione sulla mappa proiettata, corrispondente ad un pixel sulla mappa piana. Ritroviamo, quindi, in piano, un reticolo regolare equispaziato. Sottolineiamo ancora una volta che si tratta di una scelta che riguarda esclusivamente la rappresentazione sul piano dei risultati. L'algoritmo che utilizziamo per la classificazione dei dati è stato adattato (tramite l'utilizzo della distanza geodetica e la scelta di estrarre i centri in maniera uniforma sulla superficie sferica) per effettuare la classificazione sulla superficie terrestre stessa, e non sul reticolo piano rappresentato in Figura 4.11.

Una conseguenza dell'utilizzo, per la visualizzazione dei risultati, della proiezione equirettangolare, è la distorsione delle aree sulla mappa. In particolare, un'area sulla superficie terrestre posta ai poli, di uguali dimensioni rispetto ad un'altra area posta all'equatore, se proiettata su una superficie piana utilizzando 4.12 diventa molto più grande. A dimostrazione empirica di questo fatto, basta osservare in Figura 4.11 le dimensioni dell'Antartide, che proiettato sul piano sembrerebbe il continente più grande.

Prima di procedere nell'analisi, sottolineiamo un fatto importante: estrarre i centri uniformemente sulla superficie terrestre è molto diverso rispetto ad estrarli uniformemente sul reticolo piano costituito da i siti che abbiamo a disposizione. Infatti, le distanze tra i siti che compongono il reticolo piano, non sono le stesse dovunque: due siti adiacenti posti vicino ai poli sono tra loro più vicini di due siti adiacenti posti sull'equatore, in termini di distanza geodetica. Scegliendo, come abbiamo fatto, di estrarre i centri uniformemente sulla superficie di una sfera, otteniamo tasselli che hanno mediamente le stesse dimensioni. Tuttavia, dato che nella proiezione piana le aree risultano deformate, tali tasselli conterranno in media un diverso numero di siti in funzione della latitudine alla quale sono posti. In particolare, i tasselli posti a latitudini elevate con-

tengono in generale più punti di quelli a latitudini vicine allo zero, poiché i siti sono più ravvicinati nelle zone polari. Questo aspetto non costituisce un problema, poiché, come abbiamo mostrato, anche l'informazione contenuta nei dati relativi ai siti varia in funzione della latitudine. In particolare, quindi, un tassello posto vicino ai poli conterrà mediamente un numero maggiore di siti, ma tali siti hanno aree di riferimento più piccole, contengono quindi meno informazione. Al contrario, un tassello posto vicino all'equatore conterrà mediamente un numero minore di siti, ma tali siti sono relativi ad aree di dimensioni maggiori, contengono quindi più informazione. Possiamo assumere, in conclusione, che la quantità di informazione all'interno di un tassello dipenda dall'area effettiva del tassello sulla superficie terrestre, e non dal numero di siti in esso contenuti, e che quindi sia la stessa, in media, al polo e all'equatore se si sceglie, come abbiamo fatto, di estrarre i siti uniformemente sulla superficie terrestre.

Nelle Figure 4.12 e 4.13 mostriamo l'effetto della riduzione del dataset originale dopo aver generato la tassellazione e calcolato i dati di sintesi: i dati originali, ottenuti dopo aver effettuato lo smoothing sul dataset a nostra disposizione sono mostrati in Figura 4.12. Come per i dati sintetici che avevamo ottenuto con le simulazioni nel Capitolo 3, anche in questo caso le curve mostrate nel grafico sono troppe e troppo variabili per poter distinguere un andamento generale.

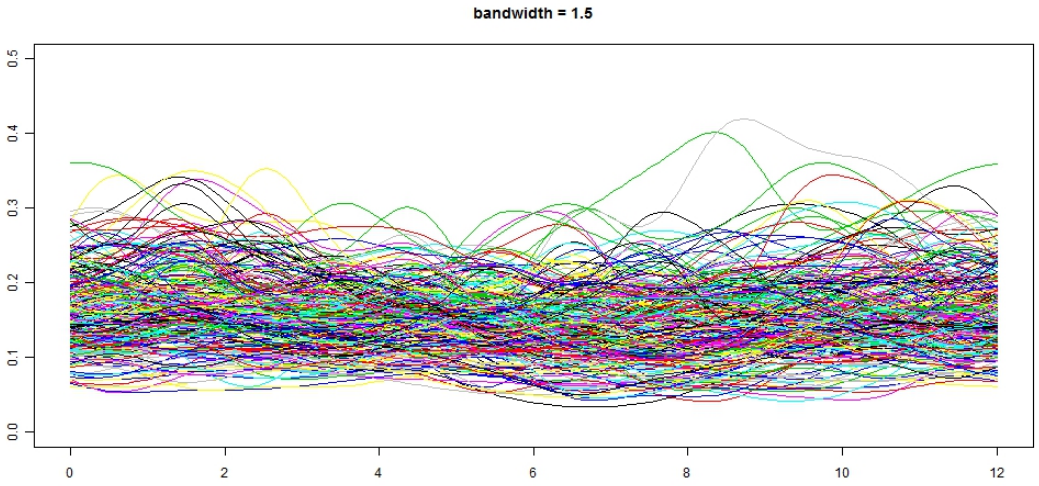


Figura 4.12: 300 dati estratti casualmente dal dataset di origine ottenuto dopo lo smoothing con kernel gaussiano, $h = 1.5$

In Figura 4.13, invece, mostriamo i dati di sintesi ottenuti dopo la tassellazione e riduzione del dataset. Sebbene nemmeno in questo caso sia possibile riconoscere gruppi distinti, notiamo che la variabilità dei dati è diminuita, come anche è diminuito drasticamente il numero di dati da classificare.

4.3.2 Riduzione dimensionale dei dati

Per la riduzione dimensionale dei dati, utilizziamo la tecnica delle componenti principali funzionali, presentata nel Capitolo 2. Una volta ottenuti i dati di sintesi dei tasselli, quindi, cerchiamo la base delle componenti principali dei dati e rappresentiamo il dato tramite gli scores ad esso associati, cioè i coefficienti associati alle diverse autofunzioni.

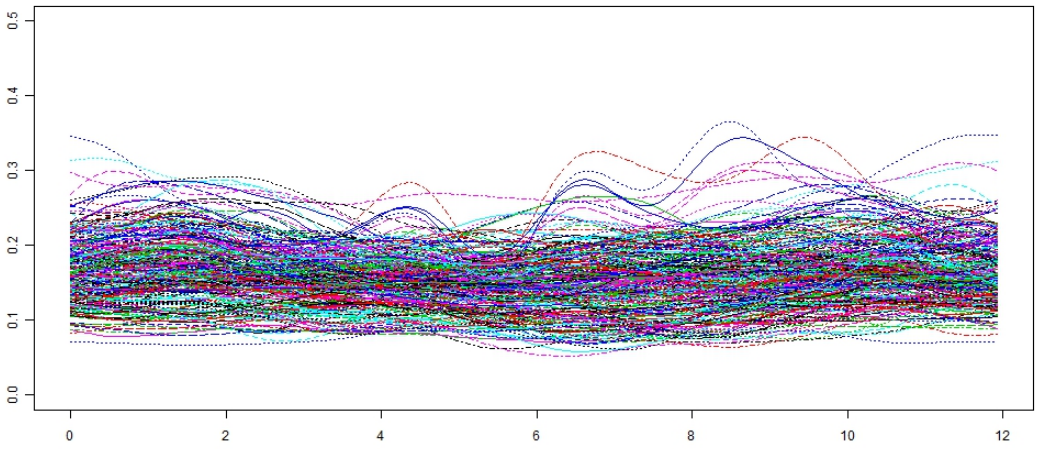


Figura 4.13: Dati di sintesi dei tasselli relativi ad una tassellazione di Voronoi con $n = 300$

Come detto, si tratta di una tecnica per ridurre la dimensione del dato, quindi arresteremo lo sviluppo ad un certo ordine q. Vediamo ora in dettaglio i risultati ottenuti per quanto riguarda la riduzione dimensionale dei dati di sintesi dei tasselli, per affrontare la scelta del numero di componenti principali da utilizzare per la classificazione.

Ricordiamo che la tecnica che abbiamo proposto prevede di effettuare riduzione dimensionale tramite FPCA sui dati di sintesi ottenuti dopo ogni diversa tassellazione di Voronoi. Ad ogni iterazione dell'algoritmo, quindi, sceglieremo una base diversa per rappresentare e classificare i dati di sintesi. Analizziamo quindi i risultati ottenuti dalla FPCA sotto un duplice punto di vista: innanzi tutto osserviamo i risultati ottenuti in una sola iterazione dell'algoritmo, al fine di scegliere il numero di componenti principali da considerare nel seguito; successivamente confronteremo tra loro le autofunzioni ottenute nelle diverse iterazioni del metodo, per analizzarne la stabilità e verificare che la scelta proposta per la dimensione dello spazio di rappresentazione dei dati di sintesi q che p stata effettuata considerando esclusivamente un'iterazione, permetta una buona rappresentazione dei dati lungo le diverse iterazioni.

Analisi di una singola iterazione

In Figura 4.14 mostriamo la base delle prime otto autofunzioni ottenute applicando la FPCA sui dati di sintesi della prima iterazione dell'algoritmo (i dati in Figura 4.13). Le linee verticali rappresentano le date corrispondenti a solstizi ed equinozi. In particolare, facendo riferimento all'emisfero boreale, le linee rossa e blu sono, rispettivamente, il solstizio d'estate e d'inverno (21 giugno e 21 dicembre). Le linee verde e arancione, invece, sono rispettivamente l'equinozio di primavera e quello d'autunno (20 marzo e 22 settembre).

Osservando la Figura, si nota la somiglianza delle autofunzioni da noi ottenute con gli elementi della base di Fourier, soprattutto per quanto riguarda le prime componenti. La prima autofunzione ha un andamento grossomodo costante; inoltre tale funzione non cambia mai segno. Le proiezioni dei dati lungo questa autofunzione, quindi, saranno

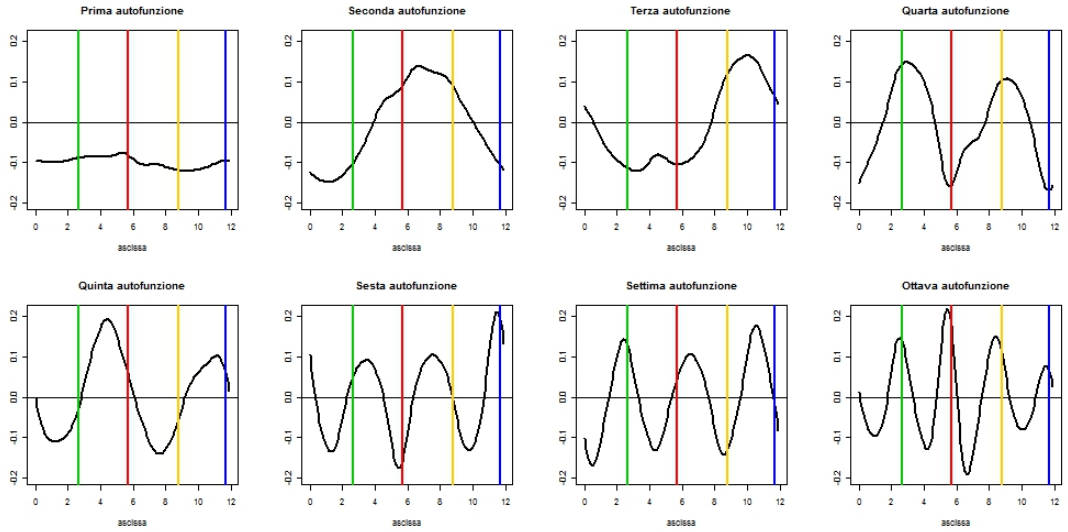


Figura 4.14: Prime otto autofunzioni ottenute tramite FPCA sul dataset composto dai dati di sintesi ottenuti dopo tassellazione di Voronoi (Figura 4.11). Linea verde=equinozio di primavera (20 marzo), linea rossa=solstizio d'estate (21 giugno), linea arancione=equinozio d'autunno (22 settembre), linea blu=solstizio d'inverno (21 dicembre)

semplicemente spostate al di sopra o al di sotto della media, a seconda degli scores associati. Dati con scores alti avranno durante tutto l'anno una proporzione di giorni senza sole più bassa della media, al contrario dati con scores bassi avranno molti giorni consecutivi senza sole rispetto alla media in tutti i periodi dell'anno.

La seconda e la terza autofunzione, invece, sono periodiche di periodo 12 mesi. Seguono un andamento abbastanza simmetrico l'una rispetto all'altra, ed entrambe hanno due cambi di segno. Evidenziano quindi il comportamento stagionale del fenomeno.

In particolare, i cambi di segno della seconda componente principale si situano nei periodi alla fine di marzo e alla fine di settembre. Questa componente, quindi, contrasta i periodi estivi e quelli invernali. Dati associati a scores positivi lungo questa autofunzione avranno periodi invernali con pochi giorni di pioggia o senza sole, e periodi estivi molto piovosi (dove invernale ed estivo è riferito a inverno ed estate australe). Dati associati a scores negativi, invece, presenteranno tale tendenza cambiata di segno, quindi inverni piovosi ed estati secche.

I cambi di segno della terza componente, invece, si situano a metà gennaio e a fine agosto. Questa componente, agisce dunque esattamente come la seconda componente principale, ma questa volta il contrasto è tra primavera ed autunno.

Proseguendo con le autofunzioni, vediamo che la frequenza dei cambi di segno aumenta progressivamente. Inoltre, la quarta e la quinta autofunzione hanno un andamento pressoché periodico di periodo 6 mesi, la sesta e la settima di periodo tre mesi e così via, analogamente alle funzioni della base di Fourier, che avevamo preso in considerazione come buona candidata per la rappresentazione dei dati.

Per quanto riguarda la scelta del numero di componenti principali da considerare nel seguito dell'analisi, osserviamo in Figura 4.15 il grafico della percentuale cumulata di variabilità totale spiegata dalle prime 10 autofunzioni. Dal grafico vediamo come, con l'utilizzo di sole 3 componenti principali, si superi ampiamente l'80% della variabilità totale.

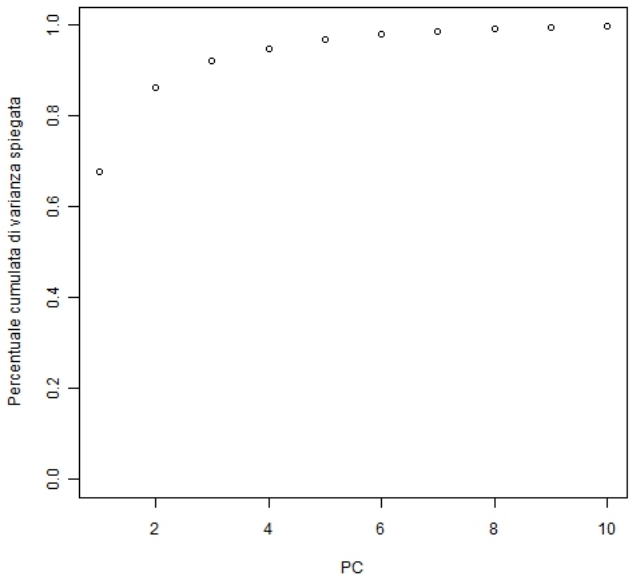


Figura 4.15: Percentuale cumulata di variabilità totale spiegata dalle prime 10 autofunzioni ottenute da FPCA dei dati di sintesi

In conclusione, i risultati mostrati nelle Figure 4.14 e 4.15, sembrano suggerire che mantenendo $q = 5$ componenti principali nell'analisi arriviamo a rappresentare completamente i dati; sembra però possibile un'ulteriore riduzione considerando esclusivamente le prime 3 componenti. Scegliamo di escludere il troncamento a $q = 4$ per mantenere il parallelo con la base di Fourier, ovvero per includere le componenti principali corrispondenti a coppie di autofunzioni tra loro simmetriche.

Confronto tra i risultati delle diverse iterazioni

Prima di procedere nell'analisi, è necessario effettuare un ulteriore controllo sulle componenti principali ottenute, per verificare che la base che identifica lo spazio nel quale rappresentiamo il dato sia stabile al variare della tassellazione di Voronoi generata ad ogni iterazione.

Innanzi tutto, in Figura 4.16 vediamo il grafico delle prime otto autofunzioni ottenute dalla FPCA di tutti i dati originali (senza quindi effettuare la riduzione del dataset tramite tassellazione di Voronoi).

Confrontando i grafici in Figura 4.16 con quelli in Figura 4.14, notiamo che le autofunzioni ottenute sono sostanzialmente le stesse. Le componenti principali dei dati che stiamo utilizzando per effettuare la classificazione, quindi, rispecchiano la variabilità dei dati originali. In particolare, non abbiamo quindi introdotto componenti fittizie attraverso la procedura di tassellazione e calcolo dei dati medi, né abbiamo eliminato o ridotto eccessivamente comportamenti originariamente esistenti.

In Figura 4.17 osserviamo invece il grafico della percentuale cumulata di variabilità spiegata dalle prime 10 autofunzioni ottenute dalla FPCA sui dati originali. Notiamo che, ora, la variabilità spiegata dalle prime componenti principali è leggermente diminuita rispetto a quanto avevamo trovato dall'analisi dei dati di sintesi. Questo perché calcolando i dati di sintesi dei tasselli abbiamo ridotto la variabilità totale del dataset.

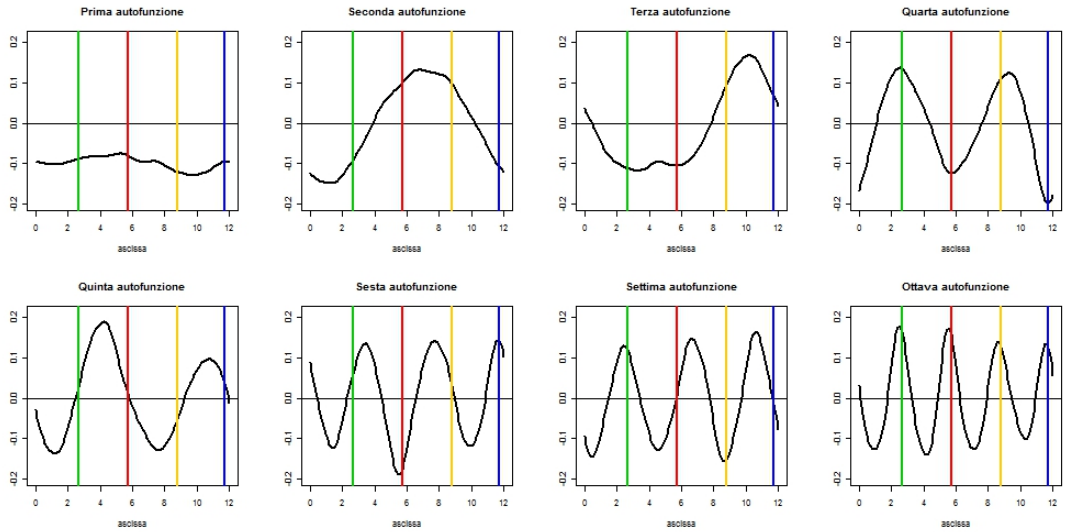


Figura 4.16: Prime otto autofunzioni ottenute tramite FPCA sul dataset composto dai dati originali. Linea verde=equinozio di primavera (20 marzo), linea rossa=solstizio d'estate (21 giugno), linea arancione=equinozio d'autunno (22 settembre), linea blu=solstizio d'inverno (21 dicembre)

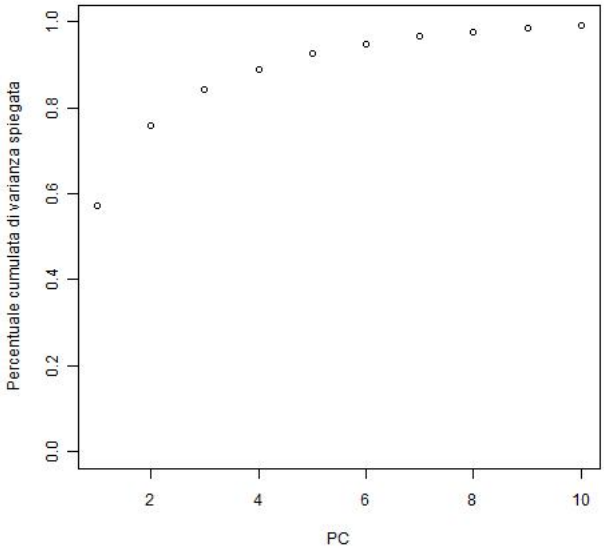


Figura 4.17: Percentuale cumulata di variabilità totale spiegata dalle prime 10 autofunzioni ottenute da FPCA dei dati originali

Una volta constatato che le prime componenti principali ottenute in una singola iterazione del metodo di classificazione proposto sono coerenti con quelle che si ottengono dalla FPCA del dataset originale, resta da verificare che tali componenti siano stabili anche su diverse iterazioni del metodo.

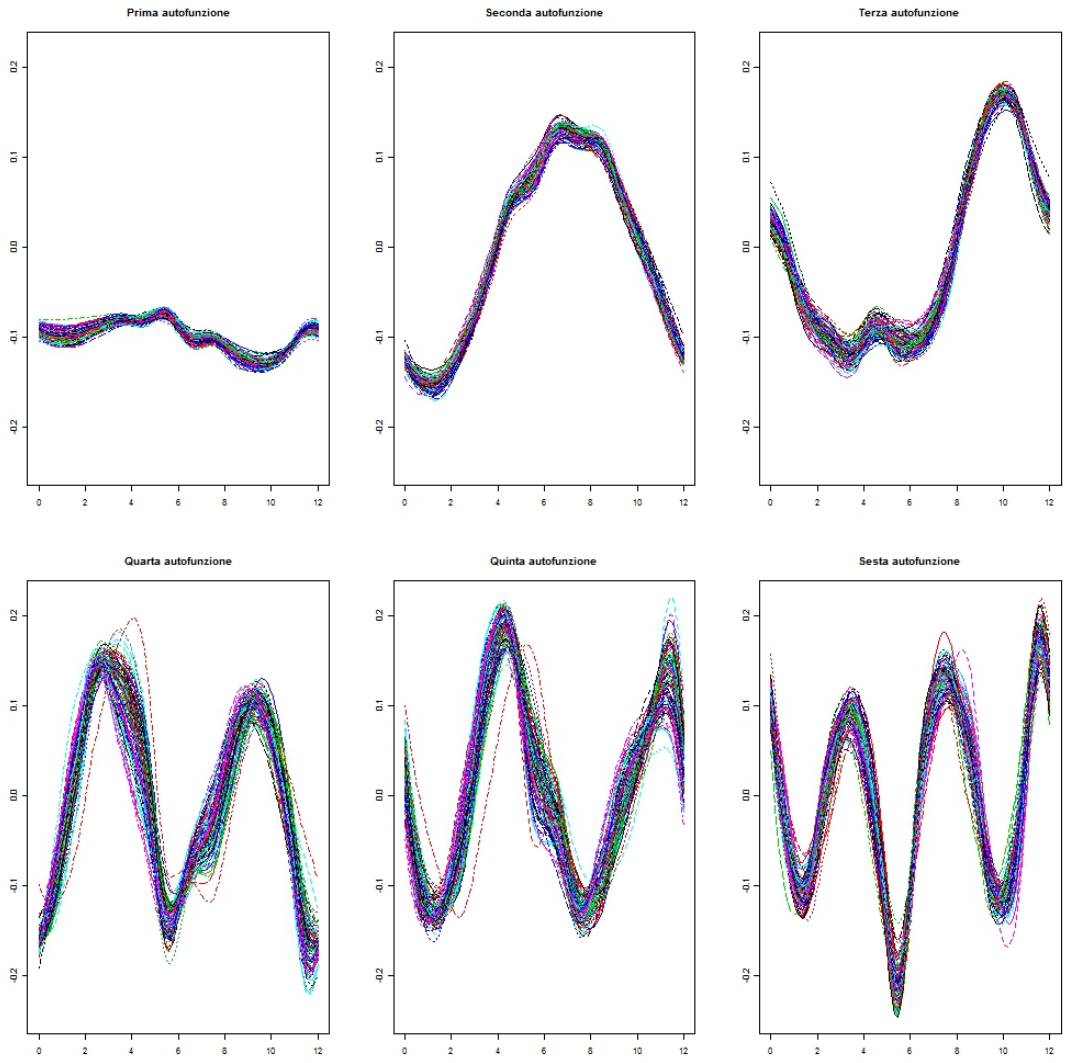


Figura 4.18: Prime sei autofunzioni ottenute tramite FPCA sul dataset composto dai dati di sintesi di 100 iterazioni del metodo proposto

In Figura 4.18 mostriamo i grafici delle prime sei autofunzioni ottenute sui dati di sintesi di 100 iterazioni della procedura proposta. Osserviamo che le autofunzioni ottenute sono molto stabili tra un'iterazione e l'altra. Sembra poi che al crescere dell'ordine delle componenti principali (quindi al diminuire degli autovalori associati, ovvero al diminuire della porzione di variabilità spiegata dalla componente), la variabilità delle atofunzioni cresca leggermente. Infatti, a partire dalla quarta componente principale, si può notare che alcune autofunzioni si discostano dall'andamento generale. Quest'ultima osservazione fornisce una ragione in più per considerare spazi di dimensione ridotta, nei quali le componenti principali ottenute sono molto stabili. In particolare, nel nostro caso, questa osservazione avvalora la scelta $q = 3$.

Per le ragioni elencate in questi paragrafi, scegliamo quindi, per proseguire l'analisi, di considerare ad ogni iterazione esclusivamente le prime tre componenti principali. Nel Paragrafo 4.3.4, poi, riassumeremo i risultati ottenuti dallo stesso metodo scegliendo $q = 5$, confrontando le due classificazioni ottenute in termini di indice di entropia e risultato finale ottenuto, per poter analizzare la robustezza dei risultati al variare del numero di componenti principali considerate.

4.3.3 Classificazione dei dati

Per l'analisi dei dati climatici che vogliamo effettuare, scegliamo in un primo momento di classificare i vettori degli scores ottenuti dalla FPCA tramite k–medie, ovvero di mantenere lo stesso metodo utilizzato per le simulazioni del Capitolo 3. Nella Sezione 4.4 presenteremo invece i risultati di classificazione ottenuti utilizzando un metodo di clustering gerarchico, per analizzare la robustezza dei risultati al variare del metodo utilizzato.

Effettueremo anche una classificazione non spaziale dei dati originali (simple clustering), per confrontare i risultati finali ottenuti con i diversi metodi. Anche la classificazione dei dati originali verrà effettuata previa riduzione dimensionale tramite FPCA e classificazione dei vettori degli scores, tramite k–medie o clustering gerarchico.

Anche in questo caso è necessario analizzare i risultati della classificazione secondo due passi successivi: inizialmente analizziamo i risultati di una singola classificazione dei dati di sintesi a partire da una particolare tassellazione di Voronoi; in questa fase analizzeremo i cluster ottenuti da una singola iterazione del metodo, osservando le curve appartenenti ai diversi cluster e la rappresentazione di tali curve nello spazio degli scores delle prime 3 componenti principali, per poi, infine, mostrare le mappe ottenute dopo una singola iterazione dell'algoritmo. Successivamente analizziamo i risultati finali della classificazione, cioè la classificazione finale stimata tramite voto di maggioranza, confrontandone visivamente i risultati con quelli ottenuti tramite simple clustering.

Ricordiamo, prima di analizzare i risultati ottenuti, che resta da fissare un ultimo parametro per poter effettuare la classificazione: il numero di cluster k, diversamente dai casi che abbiamo simulato nel Capitolo 3, non è noto. Scegliamo quindi di effettuare la classificazione per diversi valori di k (in particolare utilizzeremo $k \in \{2, 3, ..., 10\}$), per poi scegliere la classificazione migliore.

Analisi di una singola iterazione

In questa prima fase analizziamo i risultati ottenuti dalla classificazione dei dati di sintesi relativi ad una singola iterazione del metodo. Come abbiamo osservato nel Capitolo 2, tali risultati non sono da considerarsi significativi per quanto riguarda l'effettiva classificazione del dataset. Infatti, per ottenere una stima finale della classificazione, è necessario iterare la generazione casuale della tassellazione di Voronoi e assegnare ogni sito ad un cluster tramite voto di maggioranza. Nel caso che stiamo analizzando, tuttavia, è interessante analizzare anche i risultati intermedi per diverse ragioni: innanzi tutto, si tratta di dati reali, relativi alla descrizione di un fenomeno climatico non banale. Vogliamo quindi verificare che la classificazione effettuata in una singola iterazione dei dati di sintesi dei tasselli sia significativa, che colga cioè reali differenze nell'evoluzione annuale dei dati di sintesi relativi alla particolare tassellazione di Voronoi utilizzata. Inoltre, data l'elevata dimensione e variabilità del dataset, è possibile osservare gli andamenti annuali dei dati funzionali assegnati ai diversi cluster nel corso di una singola

iterazione, e tale andamento sarebbe molto difficile da osservare sul dataset originario per ogni sito, una volta calcolata la stima finale con voto di maggioranza.

Sottolineiamo ancora una volta che le analisi dei risultati presentate in questo paragrafo, relative ad una singola iterazione, sono da considerarsi significative solo ai fini dell'illustrazione dell'algoritmo sviluppato, non si tratta quindi di risultati definitivi.

Fissiamo, quindi, una particolare tassellazione di Voronoi ottenuta a partore da $n = 300$ centri, ed analizziamo i risultati della classificazione dei dati di sintesi, cioè i dati rappresentati in Figura 4.13. Una volta effettuata la riduzione dimensionale del dataset tramite FPCA, descritta nel Paragrafo 4.3.2, il metodo proposto procede effettuando la classificazione dei vettori degli scores con tecniche di analisi multivariata. Infatti, abbiamo visto nel Capitolo 2 che la distanza euclidea tra i vettori degli scores delle prime q componenti principali è una seminorma in L^2. Quindi, effettuare una classificazione multivariata dei vettori degli scores equivale ad effettuare una classificazione funzionale in base alla seminorma indotta dalle prime q autofunzioni.

Fissiamo quindi $q = 3$ la dimensione dei vettori degli scores (consideriamo, come già detto, le prime 3 componenti principali) ed effettuiamo la classificazione di tali vettori utilizzando il metodo delle k–medie.

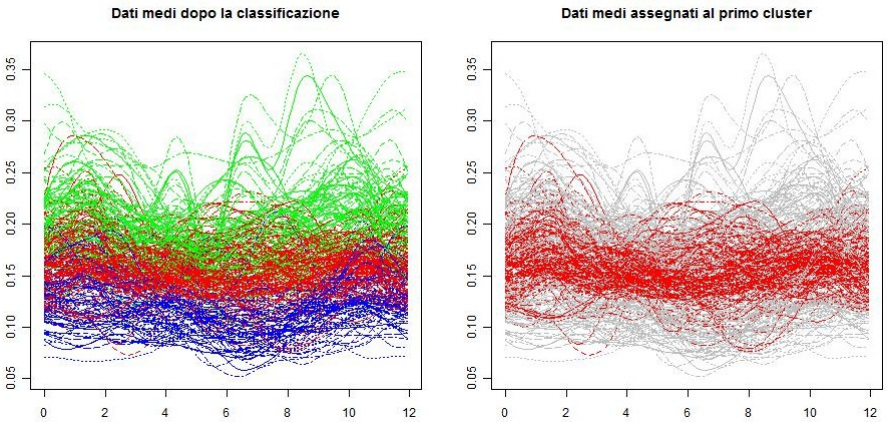
Ricordiamo infine che il numero di cluster non è fissato. Analizziamo in questo paragrafo i diversi risultati ottenuti per $k = 3, 6, 9$. Successivamente, nel Paragrafo 4.3.4, effettueremo la scelta del numero di cluster da considerare analizzando l'entropia associata alla classificazione finale.

Il risultato della classificazione di tali dati di sintesi, per $k = 3$ è mostrato in Figura 4.19, mentre per $k = 6$ nelle Figure 4.20 e 4.21. Per $k = 9$ mostriamo in Figura 4.22 la classificazione di tutte le curve, e analogamente in Figura 4.23 le curve associate ai singoli cluster.

Confrontando i grafici delle classificazioni dei dati, osservare che per $k = 3$ la classificazione viene effettuata quasi esclusivamente secondo la media delle curve, cioè in base alla prima componente principale. Infatti i tre gruppi di curve si separano esclusivamente perché hanno ordinata media diversa: si distingue un primo cluster composto dalle curve che stanno a metà rispetto al gruppo completo (zone con periodi di copertura nuvolosa di lunghezza intermedia), un secondo cluster composto dalle curve con valori più alti (quindi zone con lunghi periodi di copertura nuvolosa) ed un terzo cluster composto dalle curve con valori più bassi (quindi zone molto secche, con brevi periodi di copertura nuvolosa).

Aumentando il numero di cluster k, invece, notiamo che anche la seconda componente principale inizia a giocare un ruolo importante nella classificazione: infatti, osservando le curve relative ai diversi cluster per $k = 6$ (Figura 4.21), possiamo notare che le curve associate al primo ed al secondo cluster hanno circa lo stesso valore medio, ma mentre le curve del primo cluster hanno un andamento abbastanza costante durante l'anno, quelle associate al secondo cluster hanno una stagionalità forte. Infatti la proporzione di giorni senza sole durante l'estate boreale è molto più bassa di quella durante l'inverno boreale. Lo stesso comportamento descritto si può osservare anche per le curve relative al terzo e al sesto cluster: il terzo cluster è composto da dati stagionali mentre il sesto da dati il cui andamento annuale è pressoché costante.

Un'altra osservazione che è possibile fare da quanto detto è che sembra, dai risultati intermedi ottenuti, che la stagionalità agisca nella sola direzione specificata. Infatti, ci sono cluster nei quali la proporzione di giorni senza sole durante l'estate boreale è più bassa di quella durante l'inverno boreale, ma non si osservano cluster che presentano



(a) Dati classificati (b) Primo cluster

(c) Secondo cluster (d) Terzo cluster

Figura 4.19: Dati di sintesi ottenuti dopo tassellazione di Voronoi, classificati tramite 3–medie sugli scores

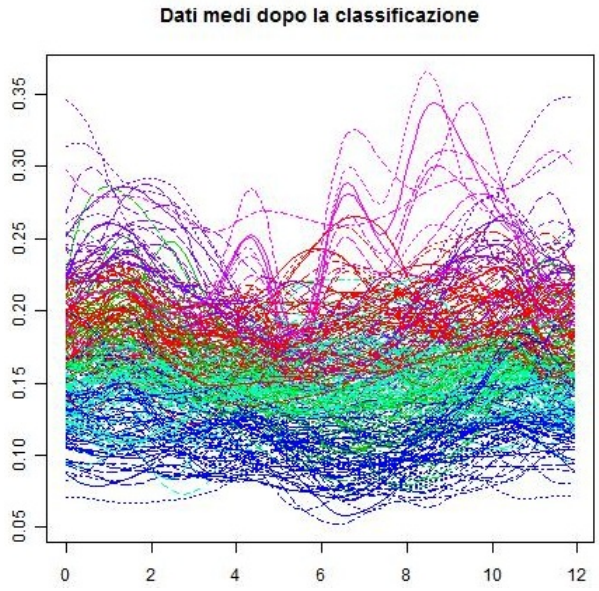


Figura 4.20: Dati di sintesi ottenuti dopo tassellazione di Voronoi, classificati tramite 6–medie sugli scores

la tendenza opposta, nei quali cioè la proporzione dei giorni consecutivi senza sole è maggiore durante l'estate boreale e minore durante l'inverno boreale.

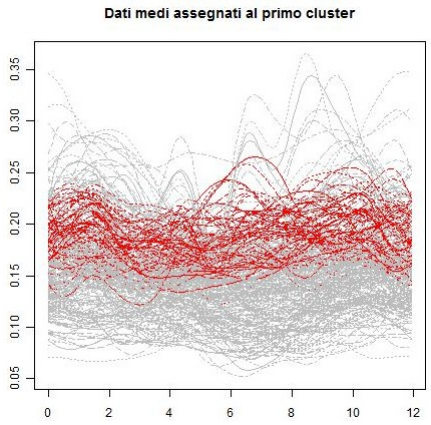
Questa osservazione è confermata dalle Figure 4.24, 4.26 e 4.26, che mostrano, rispettivamente per $k = 3$, $k = 6$ e $k = 9$, gli scatterplot degli scores associati alle prime 3 componenti principali, nelle quali i colori rispecchiano la classificazione dei dati di sintesi che abbiamo visto nelle Figure 4.19, 4.20, 4.21, 4.22 e 4.23.

Innanzi tutto notiamo che non sembra esserci un'evidente separazione in cluster negli scores. È ragionevole pensare, in effetti, che il fenomeno che stiamo osservando vari in maniera continua tra una zona e l'altra, e non ci sia una distinzione netta tra comportamenti differenti. Questa osservazione, tuttavia, non mette in dubbio la significatività dell'analisi di classificazione. Infatti, quello che stiamo facendo non è separare popolazioni tra loro diverse, ma raggruppare dati simili in modo da descrivere l'evoluzione del fenomeno osservato in maniera esauriente senza dover utilizzare tutti i dati.

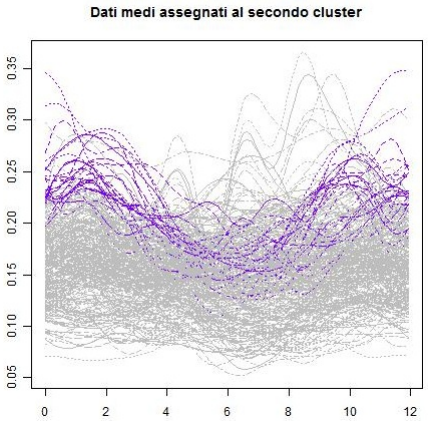
L'analogo, nell'analisi di dati scalari, di questo procedimento è la segmentazione di immagini. Un'immagine può essere infatti rappresentata da un reticolo piano regolare nel quale ogni sito rappresenta un pixel, per ognuno dei quali registriamo un valore numerico che rappresenta l'intensità del pixel associato. L'obiettivo della segmentazione di un'immagine è la suddivisione dell'immagine stessa in aree, cioè insiemi di pixel o siti, con caratteristiche omogenee al fine di semplificare l'immagine per ottenerne una rappresentazione più nitida.

Tornando a osservare i risultati della classificazione, vediamo in Figura 4.24 che per $k = 3$ la classificazione nei tre gruppi viene fatta tenendo conto esclusivamente della prima componente del vettore degli scores, cioè la prima componente principale, coerentemente con quanto avevamo osservato dai grafici della classificazione dei dati di sintesi.

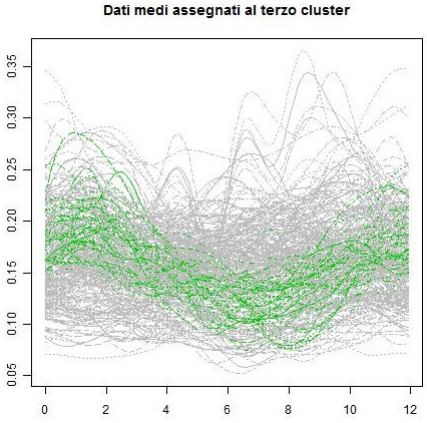
Aumentando il valore di k, invece, notiamo che anche la seconda componente principale entra in gioco nella classificazione, mentre l'apporto della terza componente resta



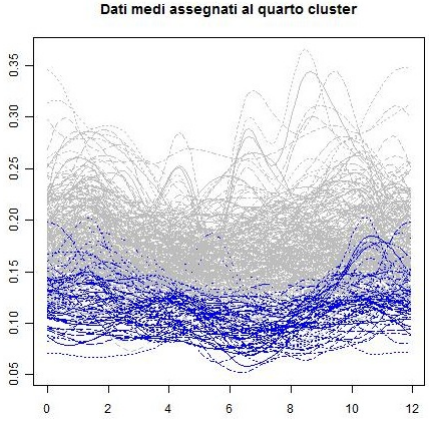
(a) Primo cluster



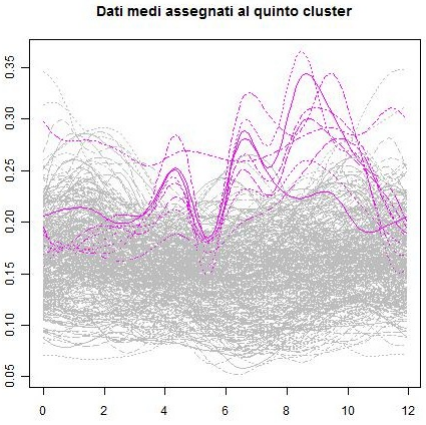
(b) Secondo cluster



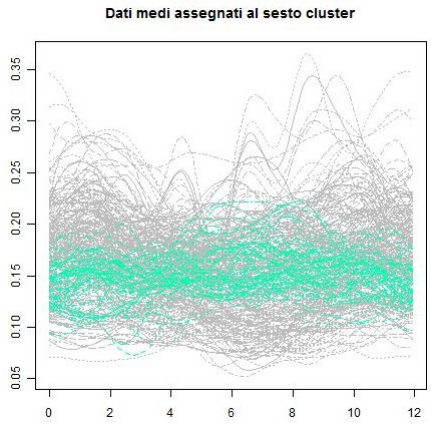
(c) Terzo cluster



(d) Quarto cluster



(e) Quinto cluster



(f) Sesto cluster

Figura 4.21: Dati di sintesi ottenuti dopo tassellazione di Voronoi, classificati tramite 6–medie sugli scores, singoli cluster

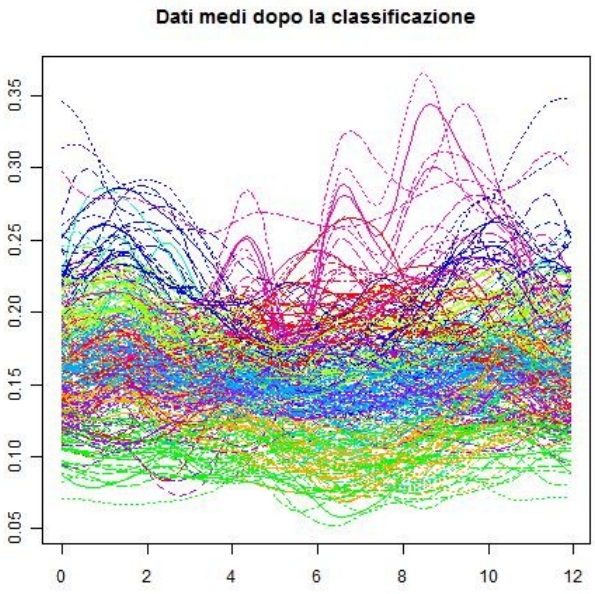


Figura 4.22: Dati di sintesi ottenuti dopo tassellazione di Voronoi, classificati tramite 9–medie sugli scores

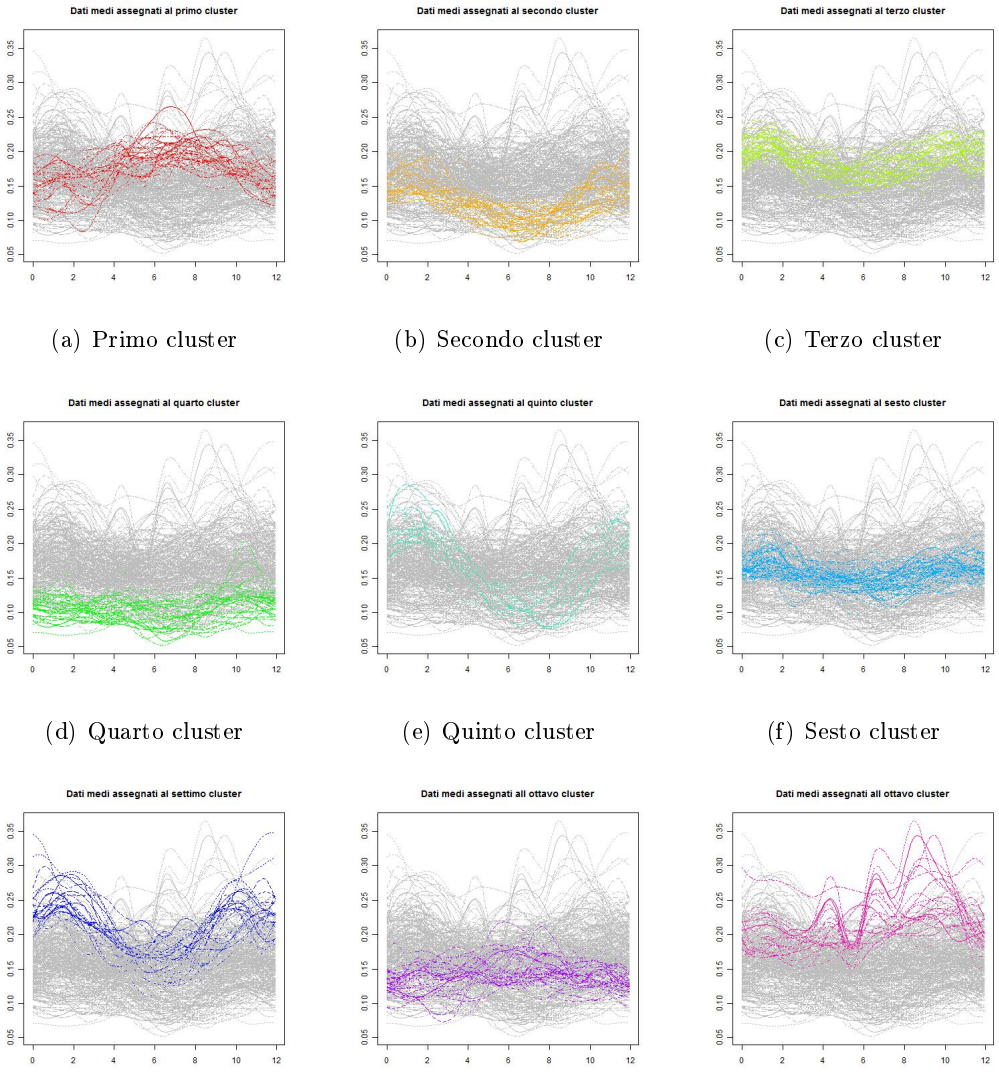
irrilevante, anche nel caso $k = 9$. Queste osservazioni giustificano la scelta di mantenere esclusivamente le prime tre componenti principali nell'analisi.

Confrontando i grafici delle Figure 4.24, 4.25 e 4.26, notiamo poi che la suddivisione in cluster per $k = 6$ sembra ottenuta a partire dalla classificazione per $k = 3$, suddividendo nuovamente,in base alla seconda componente, i due cluster relativi a scores più bassi della prima componente principale. La struttura dei cluster sembra quindi annidata.

Dalle Figure 4.24, 4.25 e 4.26 osserviamo infine che c'è un dato il cui score lungo la prima componente principale è notevolmente più basso degli altri. Osservando i valori degli scores, immaginiamo che sia un particolare dato senza tendenza stagionale, quindi con un'evoluzione globalmente costante (gli scores della seconda e terza componente principale sono circa uguali a zero), che però si trova molto più in alto rispetto agli altri dati. Si tratta, in particolare, del dato di sinesi associato ad un tassello che ricopre la zona sudorientale della Cina.

In Figura 4.27 notiamo che in effetti il dato funzionale che corrisponde a quel punto (la curva in rosso) si discosta, come andamento, da tutte le altre curve del dataset. Tale dato funzionale, che rappresenta il dato di sintesi di un tassello, è un outlier nella distribuzione delle curve osservate. Tornando ad osservare i risultati della classificazione, nelle Figure 4.19, 4.20, 4.21, 4.22 e 4.23, vediamo come questo dato venga classificato sempre assieme al cluster relativo ai dati mediamente più elevati, ma che l'andamento della frazione di giorni senza sole per questo dato è diverso, in quanto non presenta la stagionalità propria degli altri dati funzionali appartenenti allo stesso cluster.

Ricordiamo che i dati che stiamo classificando sono dati di sintesi relativi ai tasselli ottenuti con un diagramma di Voronoi. La presenza di un possibile outlier può essere causata dalla presenza, nel dataset originale, di un unico dato con un andamento molto diverso rispetto agli altri (dovuto quindi probabilmente ad un errore di misura), oppure



(a) Primo cluster (b) Secondo cluster (c) Terzo cluster

(d) Quarto cluster (e) Quinto cluster (f) Sesto cluster

(g) Settimo cluster (h) Ottavo cluster (i) Nono cluster

Figura 4.23: Dati di sintesi ottenuti dopo tassellazione di Voronoi, classificati tramite 9–medie sugli scores, singoli cluster

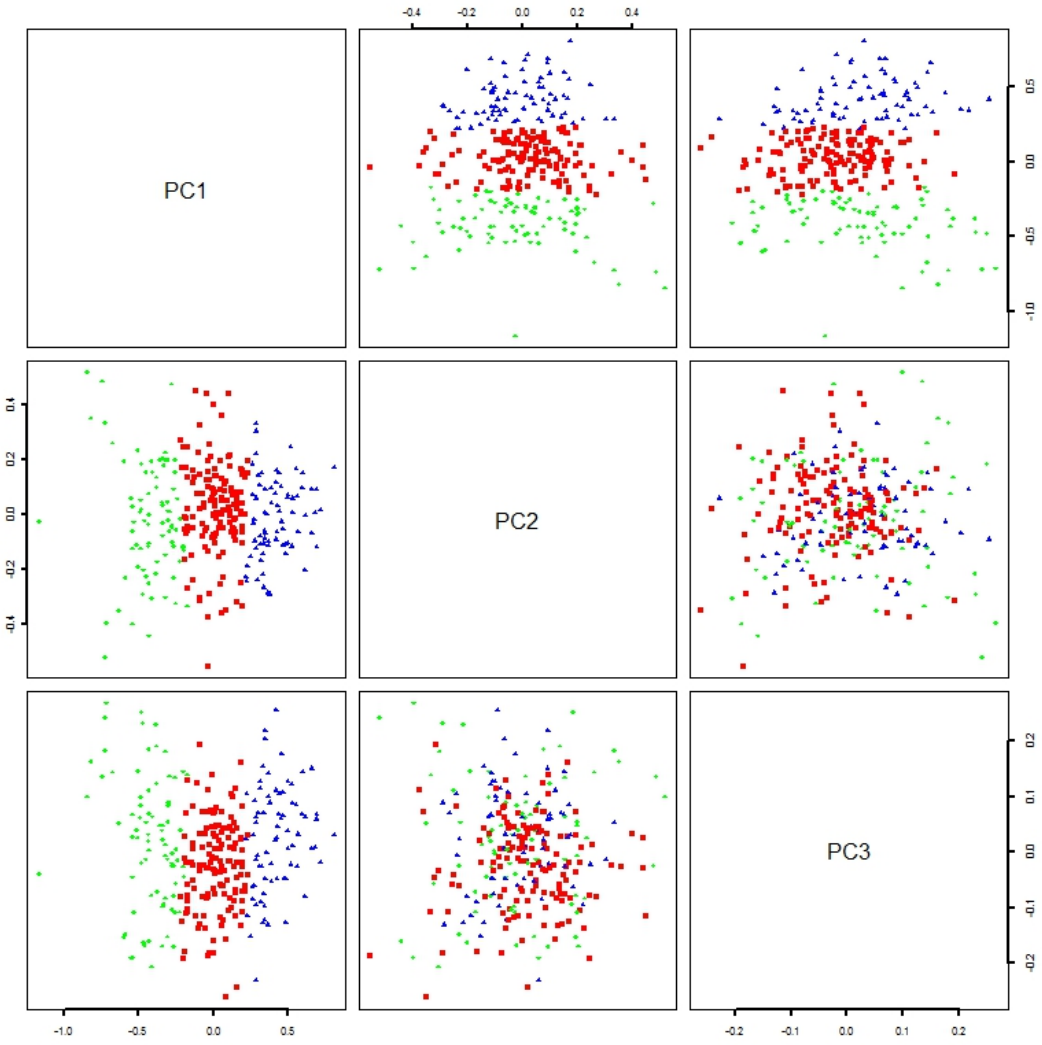


Figura 4.24: Scatterplot degli scores delle prime tre componenti principali, colorati a seconda della classificazione k–mean ottenuta con $k = 3$

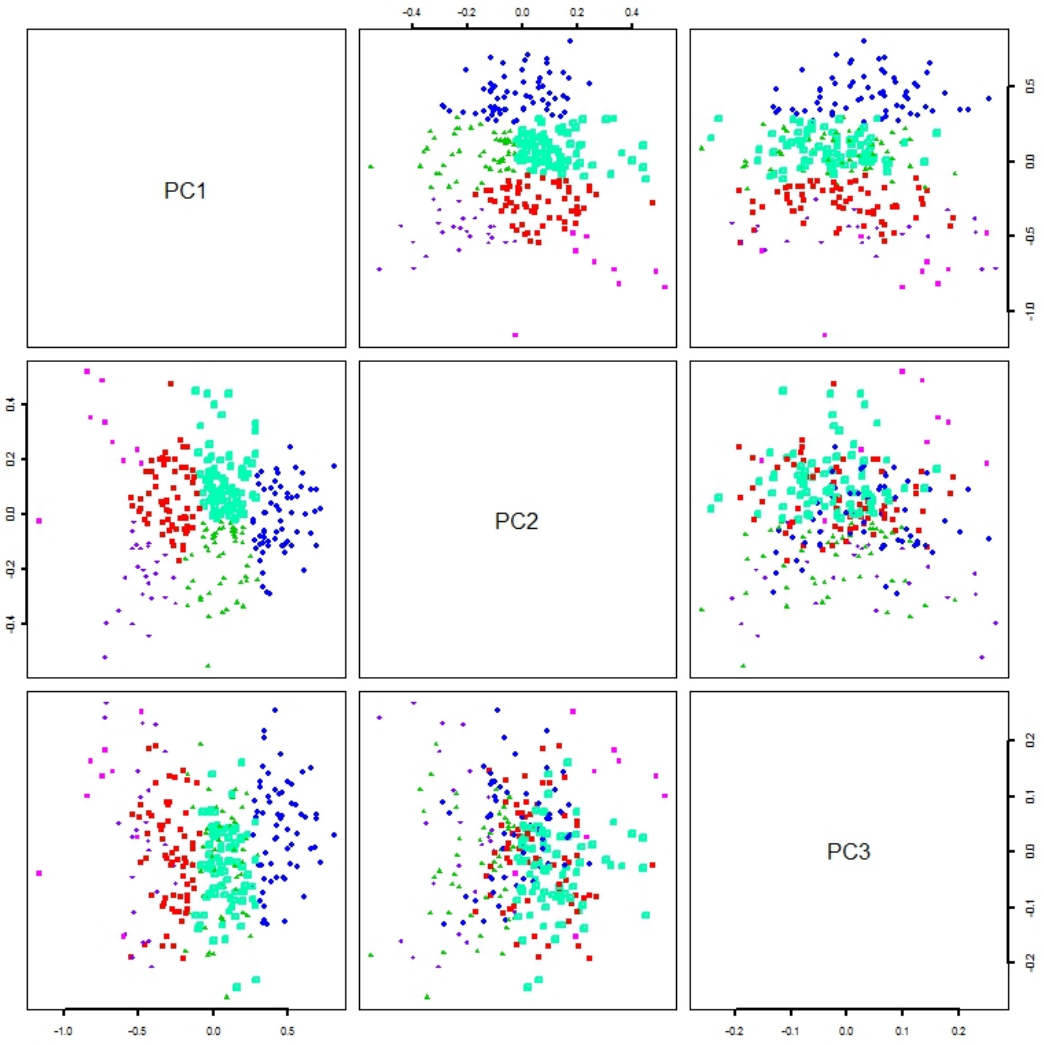


Figura 4.25: Scatterplot degli scores delle prime tre componenti principali, colorati a seconda della classificazione k–mean ottenuta con $k = 6$

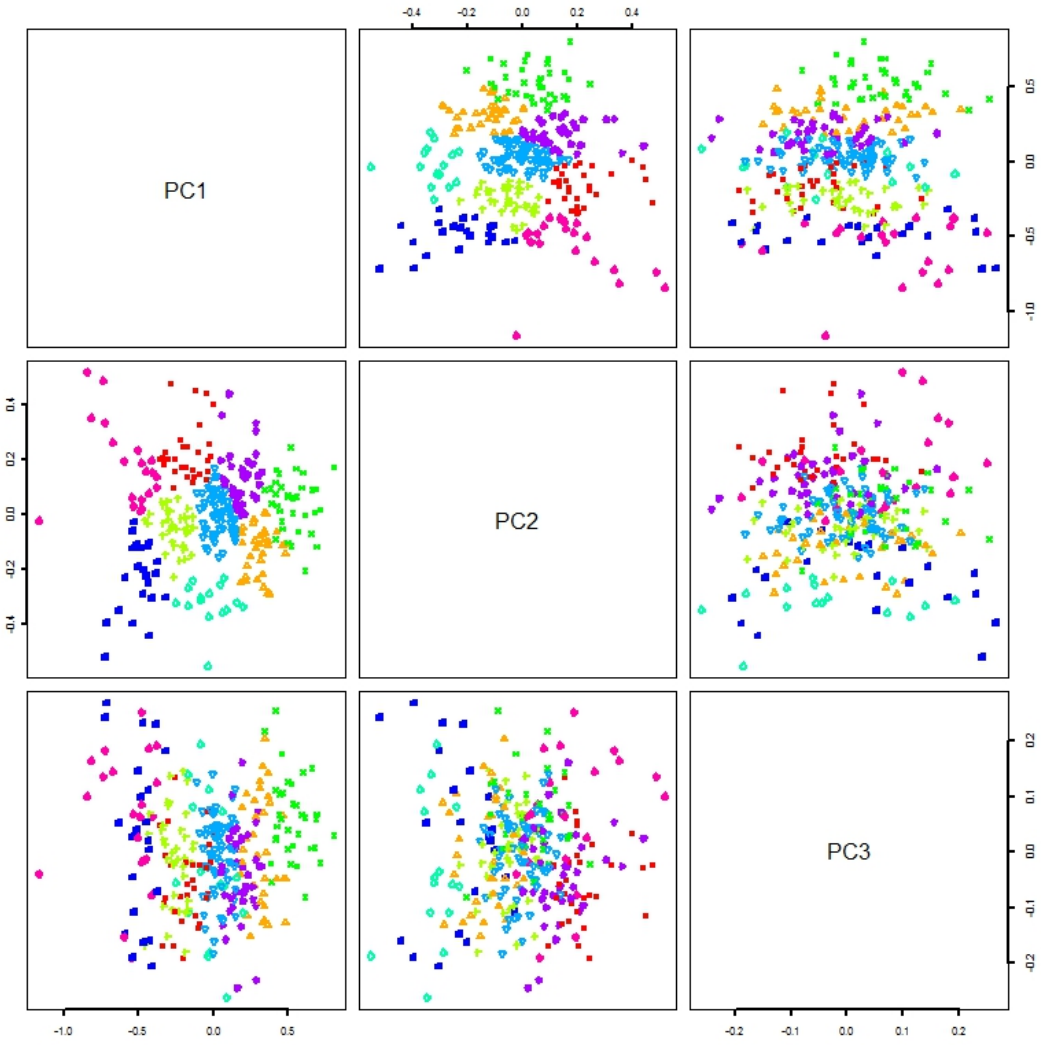


Figura 4.26: Scatterplot degli scores delle prime tre componenti principali, colorati a seconda della classificazione k–mean ottenuta con $k = 9$

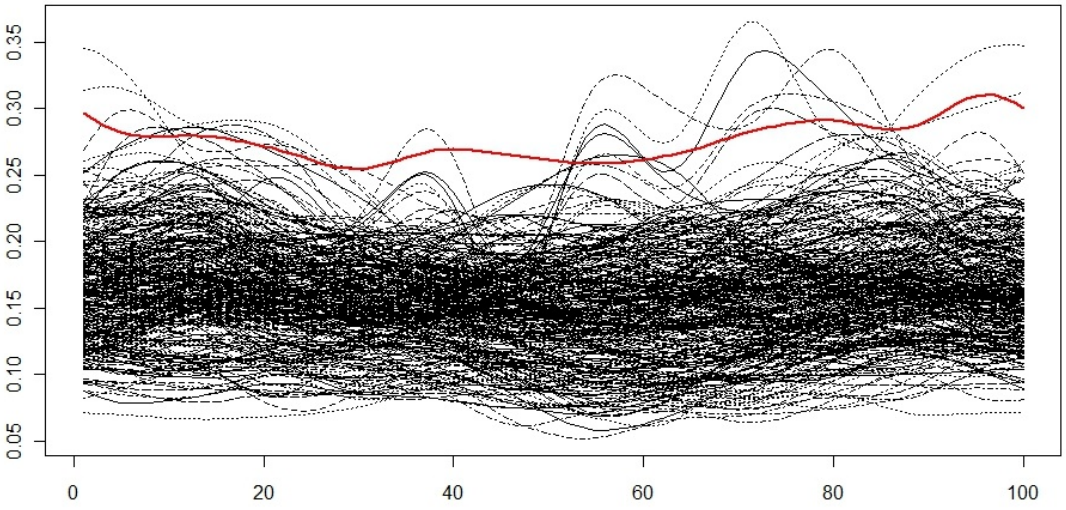


Figura 4.27: Dati di sintesi dei tasselli relativi ad una tassellazione di Voronoi con $n = 300$. In rosso è evidenziato un possibile outlier

di un'intera regione con un comportamento climatico molto diverso rispetto a quello delle regioni circostanti. Osservando esclusivamente i dati relativi a un'iterazione del metodo non ci è possibile, per ora, distinguere tra queste due situazioni. Torneremo ad analizzare il comportamento dei dati associai a siti posti in quella zona geografica in seguito, commentando le stime finali ottenute con il metodo dello spatial clustering e i risultati del simple clustering.

Osserviamo infine, nelle Figure 4.28, 4.29 e 4.30 la mappa risultante dalla classificazione dei tasselli, sempre per $k = 3, 6, 9$. Le mappe ottenute sono il risultato di una sola iterazione del metodo proposto, nelle quali di conseguenza tutti i siti appartenenti ad uno stesso tassello vengono assegnati allo stesso cluster.

Nel caso dei dati e della classificazione che abbiamo ottenuto in questa iterazione, comunque, è utile osservare i risultati delle Figure 4.28, 4.29 e 4.30, per associare visivamente la classificazione ottenuta ai tasselli. Osserviamo che innanzi tutto si ha una buona suddivisione dei diversi cluster nello spazio, e si iniziano a distinguere, soprattutto per $k = 3$ o 6 delle zone in cui suddividere la mappa. Questo risultato, conferma già in parte l'esistenza di una struttura spaziale nel dataset analizzato. Infatti, sebbene la classificazione degli scores associati ai dati di sintesi di tasslli non tenga più conto delle informazioni spaziali (che supponiamo già contenute nei dati di sintesi), si riescono comunque ad ottenere delle zone connesse nello spazio ed associate allo stesso cluster, caratteristica che non era garantita, a livello teorico, dal metodo utilizzato. Se, infatti, i dati fossero stati tra loro indipendenti, avremmo più probabilmente osservato delle immagini nelle quali le assegnazioni dei tasselli ai diversi cluster avrebbero dato origine a una disposizione apparentemente casuale dei colori sulla mappa.

In secondo luogo, dalle immagini osserviamo che il dato che rappresenta un possibile outlier, o quantomeno una zona con un andamento annuale dei periodi senza sole abbastanza particolare, è quello relativo al tassello che ricopre la zona costiera della Cina, che infatti risulta classificato insieme alle zone subpolari artiche in tutte e tre le classificazioni.

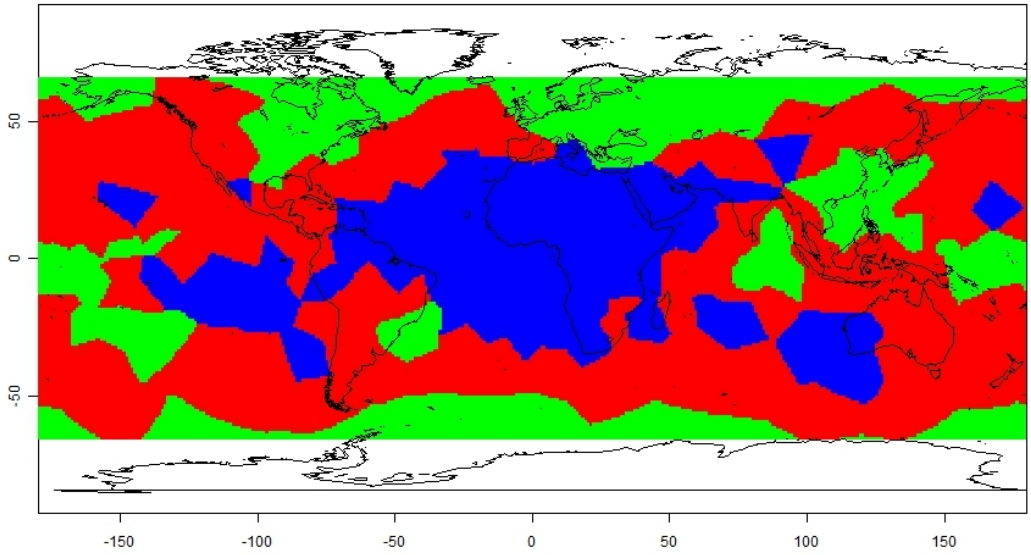


Figura 4.28: Mappa risultante dopo un'iterazione dell'algoritmo per $k = 3$

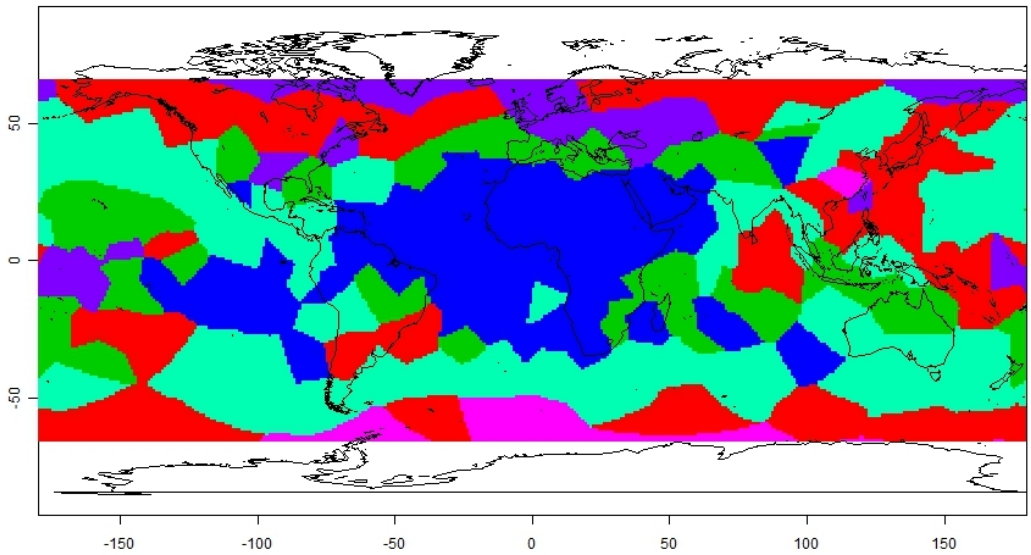


Figura 4.29: Mappa risultante dopo un'iterazione dell'algoritmo per $k = 6$

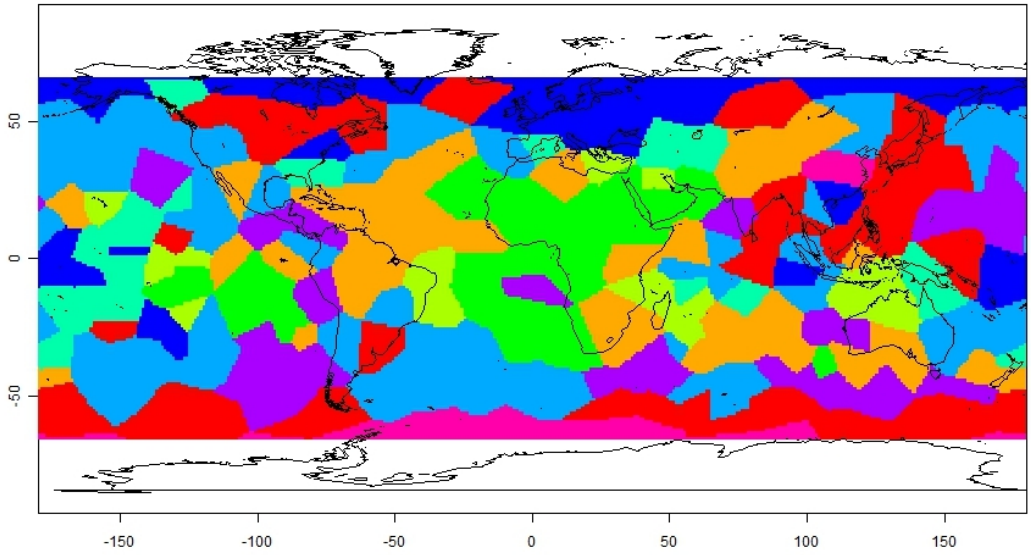


Figura 4.30: Mappa risultante dopo un'iterazione dell'algoritmo per $k = 9$

Risultati dopo $M = 100$ iterazioni

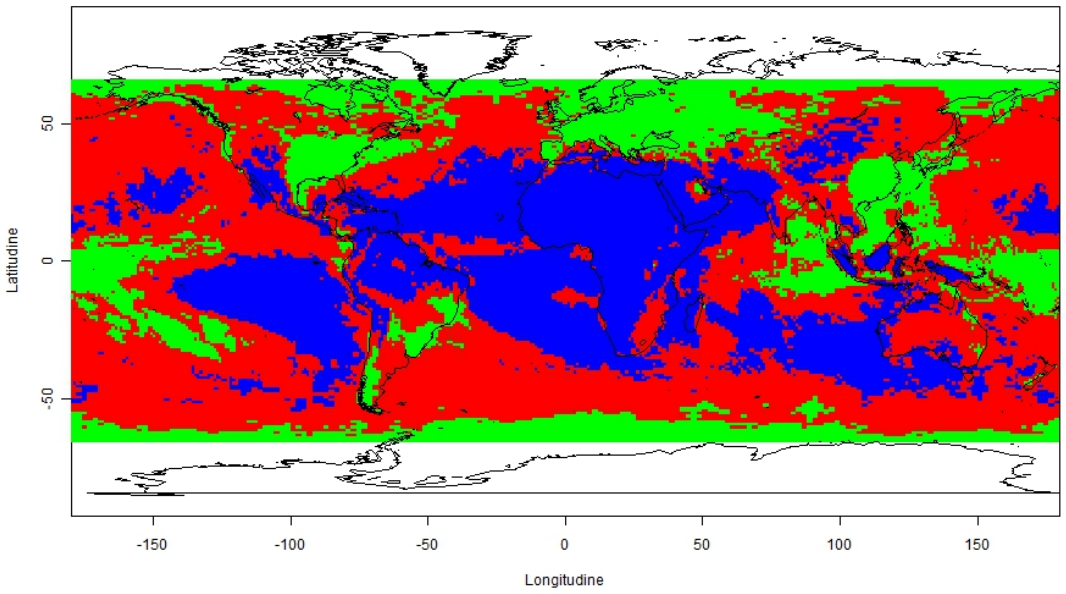
In base alla procedura di classificazione descritta nel Capitolo 2, per ottenere una stima della clusterizzazione finale è necessario iterare la generazione della tassellazione di Voronoi e la classificazione dei dati di sintesi dei tasselli, per poi assegnare ogni sito della mappa al cluster al quale è stato assegnato più volte nel corso delle diverse iterazioni, tra loro indipendenti.

Scegliamo quindi $M = 100$ ed effettuiamo la classificazione del dataset al variare del numero di cluster k. Come nel paragrafo precedente, effettuiamo la classificazione per $k \in \{2, 3, .., 10\}$, e mostriamo i risultati ottenuti per $k = 3, 6, 9$, che costituiscono un campione rappresentativo delle analisi effettuate.

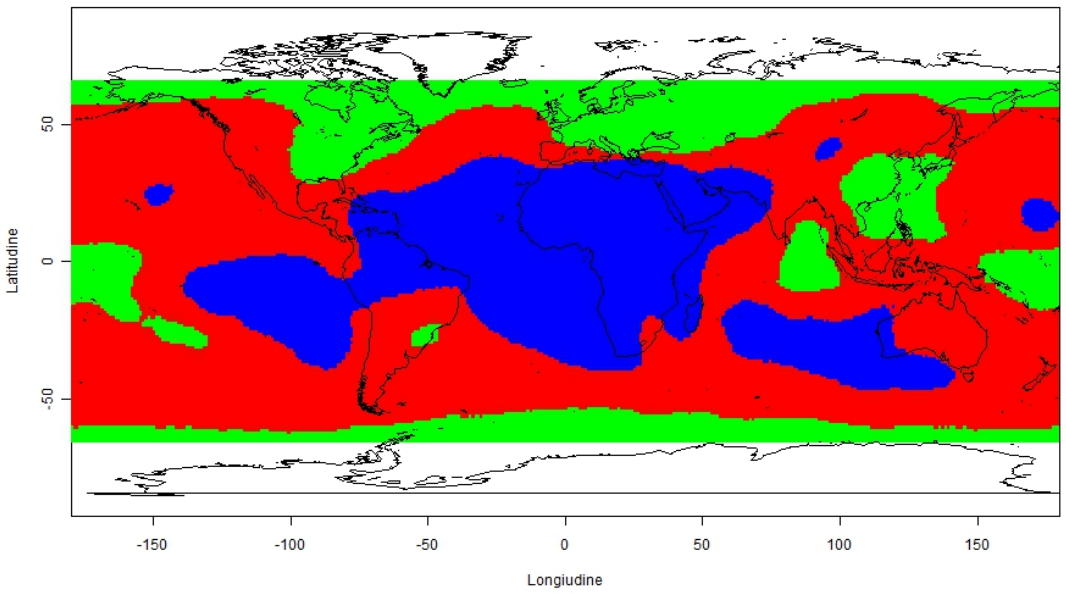
Per avere un confronto del risultato ottenuto con la tecnica di classificazione proposta rispetto a quelle standard, effettuiamo inoltre la classificazione non spaziale del dataset, applicando il metodo delle k–medie agli scores delle prime tre componenti principali ottenute dall'intero dataset originale, invece che dai dati di sintesi dei tasselli (simple clustering).

Nelle Figure 4.31, 4.32 e 4.33 mostriamo i risultati ottenuti con i due metodi, rispettivamente per $k = 3, 6$ e 9. Ovviamente, per effettuare un confronto tra i due risultati,non è possibile calcolare un rate di misclassificazione, poiché le etichette originali non sono note. Il calcolo dell'entropia, invece, può fornire un indice per scegliere tra le classificazioni ottenute con l'algoritmo di classificazione spaziale proposto al variare di k. L'entropia, però, non può essere utilizzata per scegliere tra i due metodi utilizzati, poiché nel caso della classificazione ottenuta con il metodo non spaziale, l'assegnazione ad un determinato cluster viene fatta una volta sola, e l'entropia associata alla classificazione è dunque identicamente nulla.

Non avendo a disposizione metodi quantitativi per scegliere tra due diversi metodi, ci limitiamo ad osservare visivamente il grado di nitidezza delle mappe ottenute. In tutti tre i casi, infatti, le mappe risultanti dai due metodi presentano lo stesso anda-

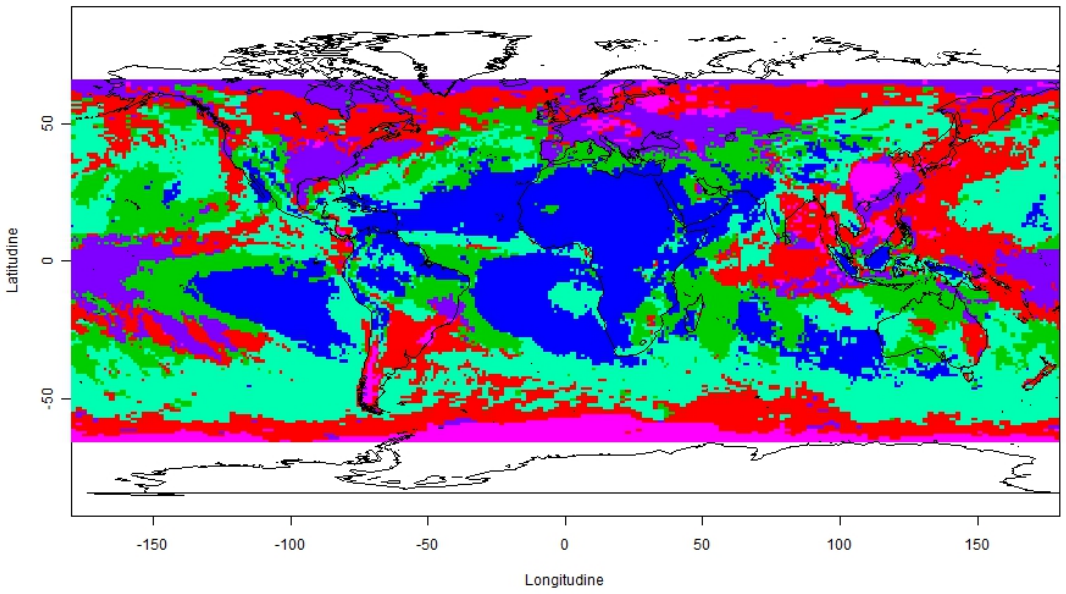


(a) Classificazione non spaziale (riduzione dimensionale e classificazione del dataset completo)

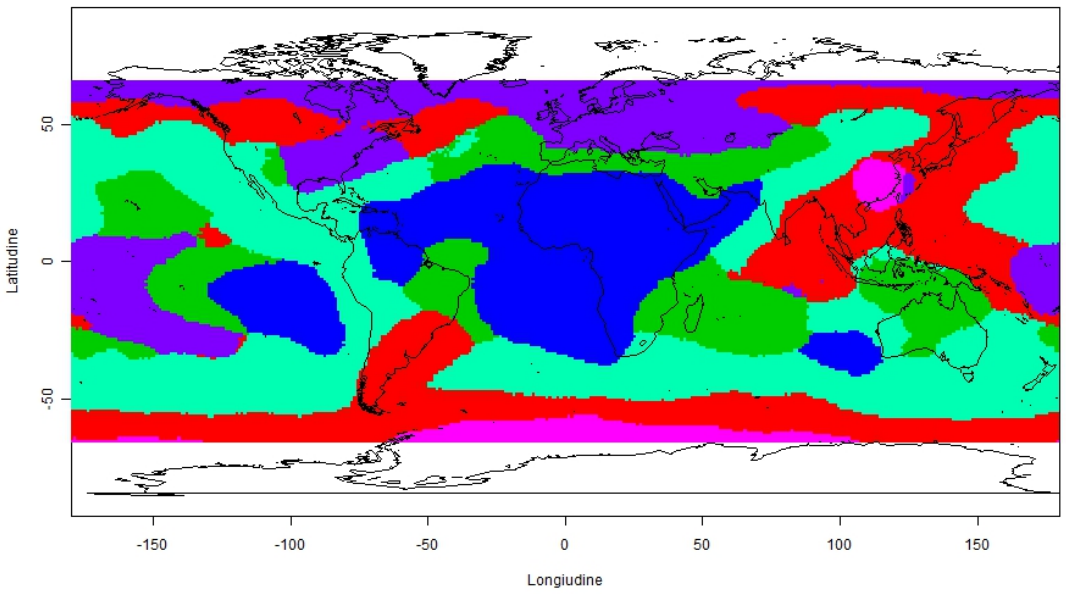


(b) Classificazione spaziale (voto di maggioranza sui risultati ottenuti nelle diverse iterazioni della procedura proposta ovvero generazione di una tassellazione di Voronoi, riduzione dimensionale, e classificazione dei dati di sintesi)

Figura 4.31: Stima finale della clusterizzazione ottenuta con k–medie non spaziale sui dati originali (a) e con la tecnica di classificazione spaziale proposta (b) per $k = 3$

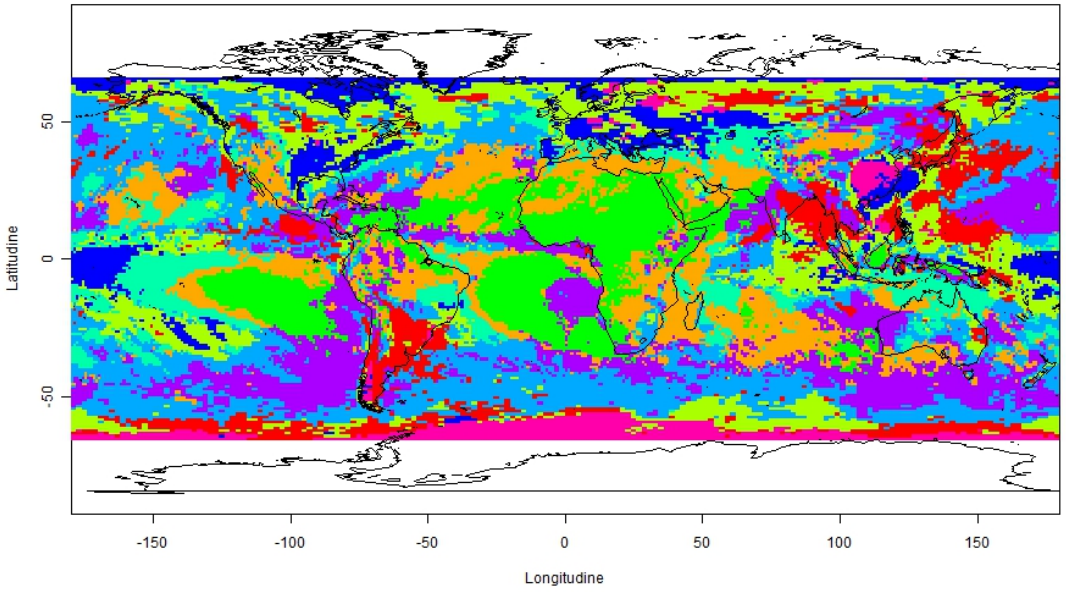


(a) Classificazione non spaziale (riduzione dimensionale e classificazione del dataset completo)

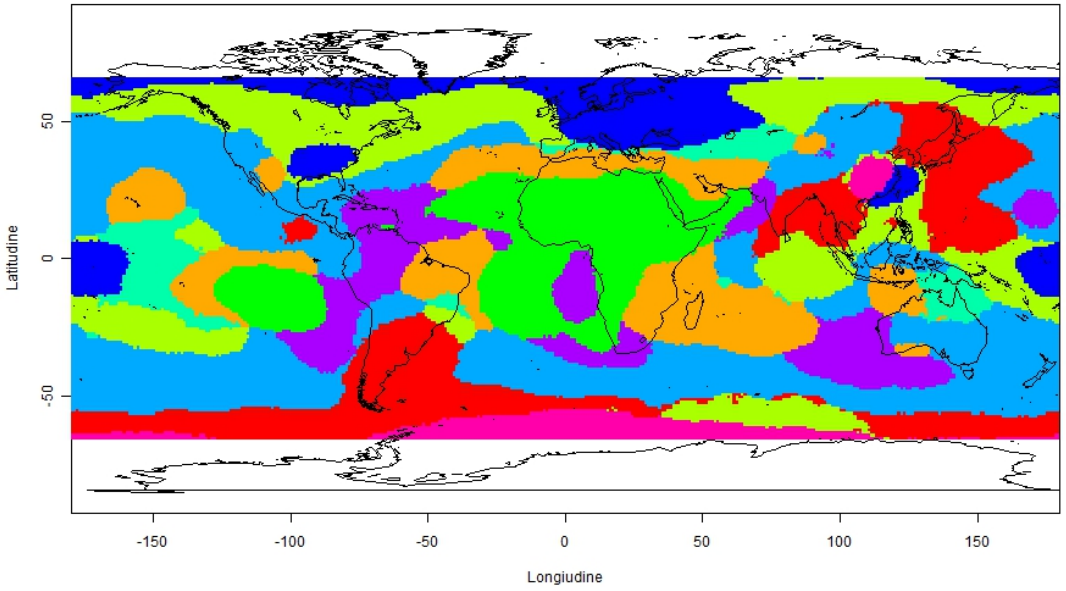


(b) Classificazione spaziale (voto di maggioranza sui risultati ottenuti nelle diverse iterazioni della procedura proposta ovvero generazione di una tassellazione di Voronoi, riduzione dimensionale, e classificazione dei dati di sintesi)

Figura 4.32: Stima finale della clusterizzazione ottenuta con k–medie non spaziale sui dati originali (a) e con la tecnica di classificazione spaziale proposta (b) per $k = 6$



(a) Classificazione non spaziale (riduzione dimensionale e classificazione del dataset completo)



(b) Classificazione spaziale (voto di maggioranza sui risultati ottenuti nelle diverse iterazioni della procedura proposta ovvero generazione di una tassellazione di Voronoi, riduzione dimensionale, e classificazione dei dati di sintesi)

Figura 4.33: Stima finale della clusterizzazione ottenuta con k–medie non spaziale sui dati originali (a) e con la tecnica di classificazione spaziale proposta (b) per $k = 9$

mento generale; mentre, però, nel risultato ottenuto tramite classificazione non spaziale (simple clustering) i confini tra i diversi cluster sono molto incerti, e ci sono moltissimi punti isolati[1], nel risultato ottenuto con la tecnica di classificazione spaziale (spatial clustering) i confini tra le regioni sono netti, e visivamente il risultato è molto più nitido e facilmente interpretabile.

Osserviamo poi che, al crescere del numero di cluster, il risultato ottenuto con il primo metodo peggiora notevolmente. Per $k = 9$ si percepisce una struttura generale, ma molti cluster non si riescono a distinguere, né tantomeno ad interpretare, proprio per la presenza di moltissimi punti isolati.

Con il metodo dello spatial clustering, invece, vediamo che la mappa risultante resta molto chiara anche all'aumentare del numero dei cluster, anche se alcune zone sono più difficilmente interpretabili, rispetto ai casi $k = 6$ e $k = 3$. Infatti, per $k = 3$ i cluster ottenuti sono interpretabili come zone a periodi di copertura nuvolosa mediamente brevi, intermedi o lunghi.

Per $k = 6$, si aggiunge il fattore stagionale nella classificazione. Abbiamo infatti ancora un cluster centrale che include le zone secche non stagionali: questo cluster segue la forma del cluster delle zone secche ottenuto per $k = 3$, anche se viene leggermente rimpicciolito. Il cluster relativo alle zone temperate, a seconda della stagionalità, viene suddiviso in due cluster più piccoli, e il cluster delle zone piovose viene diviso in tre.

Confrontando le stime finali ottenute per $k = 3$ e $k = 6$ mostrate nelle Figure 4.31 (b) e 4.32 (b), notiamo che, tuttavia, i cluster non sono esattamente annidati: quello centrale, come già accennato, per $k = 6$ è più piccolo; i due cluster che per $k = 6$ dovrebbero ricoprire le zone temperate della classificazione ottenuta per $k = 3$, si estendono anche in zone appartenenti al cluster centrale, e l'unione dei tre cluster relativi per $k = 6$ a zone con alta nuvolosità è più grande del cluster delle zone con alta nuvolosità di $k = 3$.

Osservando ancora le stesse Figure, vediamo che anche nella stima finale ottenuta con spatial clustering la costa della Cina viene classificata, per tutti i diversi valori di k, assieme alle zone subpolari. Osservando i risultati ottenuti con il simple clustering, poi, possiamo vedere che, anche classificando ogni sito singolarmente, l'intera zona viene classificata, allo stesso modo, insieme alle zone subpolari. Questa osservazione esclude l'ipotesi che esista un unico sito in quella zona per il quale le informazioni sono molto diverse rispetto agli altri siti. Si tratta piuttosto di una situazione climatica molto particolare localizzata nell'intera zona.

4.3.4 Analisi dell'entropia e scelta del numero di cluster

Una volta analizzati i risultati ottenuti con il metodo di classificazione spaziale proposto, e avendoli confrontati con la stessa classificazione effettuata con i metodi standard adatti alla classificazione dei dati in ambito non spaziale, resta da scegliere, tra i diversi risultati ottenuti, la classificazione migliore, al variare del numero di cluster k. Come proposto nel Capitolo 2, effettueremo tale scelta confrontando i risultati ottenuti in termini di entropia della classificazione; tale indice, calcolabile a partire dalle frequenze di assegnazione ai diversi cluster, fornisce infatti una misura della nitidezza della mappa ottenuta. Per il calcolo dell'entropia media, apportiamo una modifica a quanto mostrato nella Sezione 2.4. Infatti, abbiamo mostrato nel Paragrafo 4.3.1 che le aree $A_{\lambda,\theta}$ sulla superficie terrestre alle quali si riferiscono i siti hanno superfici diverse al variare della

[1] Per *punti isolati* si intendono singoli siti, o insiemi di pochi siti, che nel risultato classificazione vengono associati ad un determinato cluster, mentre la regione che li contiene viene assegnata ad un diverso cluster

latitudine. Per tenere conto di tale problema, invece che calcolare la media dell'entropia pesando allo stesso modo l'entropia di ogni sito, effettuiamo una media pesata rispetto alla superficie di riferimento del sito. In questo modo, siti che si riferiscono a superfici maggiori avranno un peso maggiore ai fini del calcolo dell'entropia media.

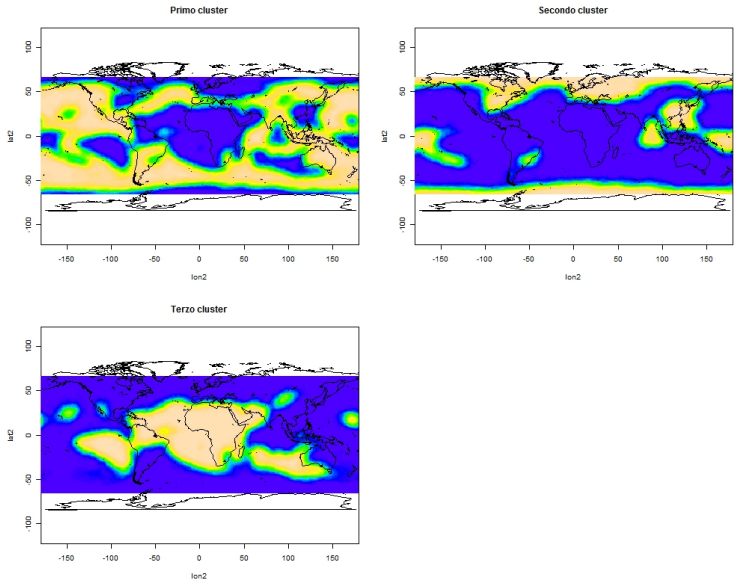


Figura 4.34: Frequenza di assegnazione dei siti della mappa ai diversi cluster lungo le iterazioni, per $k = 3$; i valori vicini a 1 sono mostrati in bianco, quelli vicino a 0 in blu

Prima di osservare direttamente i risultati dell'entropia al variare del numero di cluster, nelle Figure 4.34, 4.35 e 4.36 sono mostrate le frequenze di assegnazione dei siti ai diversi cluster, sempre per $k = 3, 6$ e 9.

La scala cromatica utilizzata, la stessa per tutti i grafici mostrati, raffigura in colore bianco i siti la cui frequenza di assegnazione ad un dato cluster è uguale a uno, in blu i siti la cui frequenza di assegnazione ad un dato cluster è zero, e in un colore intermedio (da giallo a verde) i siti con frequenze di assegnazione comprese tra zero e uno.

In base a quanto detto, una classificazione può essere considerata buona se la la maggior parte delle aree della mappa sono bianche o blu, e sono presenti poche aree dove si vedono colori intermedi, come azzurro o verde. Infatti, quando ciò accade significa che le classificazioni ottenute tra un'iterazione e l'altra del metodo sono tra loro coerenti. Visivamente si nota che ciò accade in tutti i casi analizzati: nelle mappe delle frequenze di assegnazione, infatti, la maggior parte dei cluster sono ben definiti, con aree nettamente bianche (o gialle) su fondo blu.

Ci aspettiamo, quindi, di trovare valori alti dell'entropia esclusivamente in corrispondenza delle aree di confine tra i diversi cluster, quelle zone cioè i cui siti vengono associati dalla tassellazione di Voronoi alternativamente a tasselli appartenenti a gruppi diversi, e che quindi vengono associati alternativamente a diversi cluster nel corso delle iterazioni del metodo.

Dalle stesse immagini notiamo poi che aumentando il numero di cluster, per esempio passando da 3 a 6, le aree di assegnazione incerta aumentano leggermente. La stessa osservazione si può fare passando da 6 a 9 cluster. Ci aspettiamo, quindi, di osservare un aumento dell'entropia all'aumentare del numero dei cluster. Tale aumento è, come già detto, in parte spiegabile in questo modo: aumentando il numero dei gruppi in cui

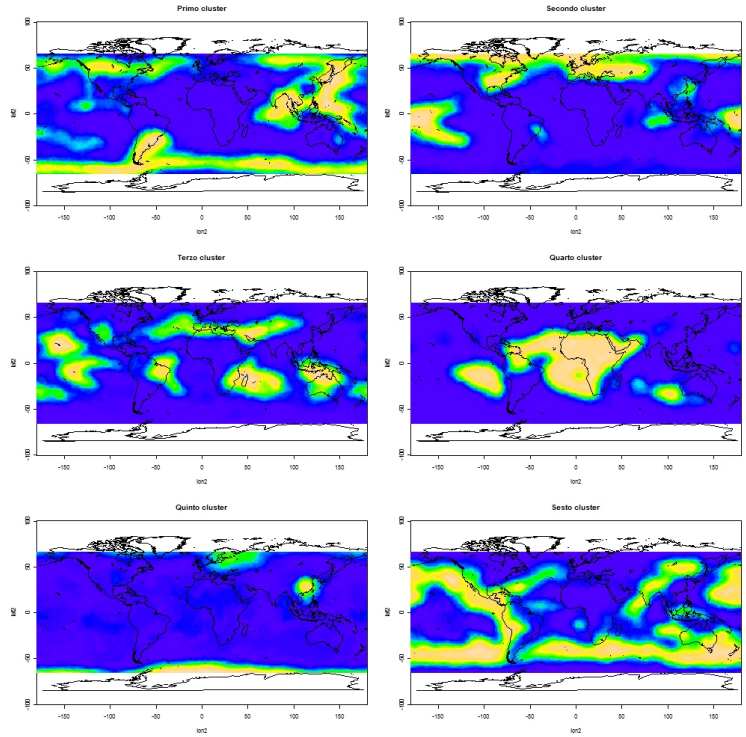


Figura 4.35: Frequenza di assegnazione dei siti della mappa ai diversi cluster lungo le iterazioni, per $k = 6$; i valori vicini a 1 sono mostrati in bianco, quelli vicino a 0 in blu

suddividere la mappa, aumentano necessariamente le aree di confine tra un cluster e l'altro, causando un aumento dell'entropia media. Analizzando l'entropia media, quindi, se osserviamo un aumento di entropia in funzione del numero di cluster dovremo capre se l'aumento è dovuto esclusivamente all'aumento delle zone di confine, o a una minore stabilità della classificazione attraverso le iterazioni.

Osserviamo quindi, nelle Figure 4.37, 4.38 e 4.39, le mappe dell'entropia delle classificazioni ottenute, sempre per $k = 3$, 6 e 9. La scala cromatica utilizzata rappresenta questa volta in rosso siti associati a valori dell'entropia bassi (vicini allo zero), in bianco siti associati a valori alti dell'entropia (vicini a uno), e in colori intermedi tra il giallo e il rosso siti associati a valori intermedi di entropia.

Dalle mappe dell'entropia mostrate, notiamo che effettivamente, come già avevamo osservato dalle frequenze di assegnazione, all'aumentare del numero di cluster si osserva un aumento dell'entropia.

La mappa dell'entropia ottenuta per $k = 3$ mostrata in Figura 4.37 si distinguono molto bene le aree di confine tra i diversi cluster, che sono effettivamente le uniche aree della mappa in cui l'entropia è alta. Osserviamo poi che le zone nelle quali l'entropia è nettamente maggiore, sono quelle in cui si osserva un contatto, o un avvicinamento, dei confini tra tre cluster, invece che tra due. Questo aspetto si osserva molto bene nella zona che comprende il mar Mediterraneo, l'Italia meridionale e la Turchia. Tornando ad osservare la stima finale per $k = 3$ in Figura 4.31, vediamo che si tratta effettivamente di una zona di confine tra tutti e tre i cluster.

Osservando la mappa dell'entropia ottenuta con $k = 6$, (Figura 4.38) si riescono ancora a distinguere nella mappa dell'entropia i confini tra i diversi cluster, anche se in modo molto meno netto rispetto a quanto osservato per $k = 3$.

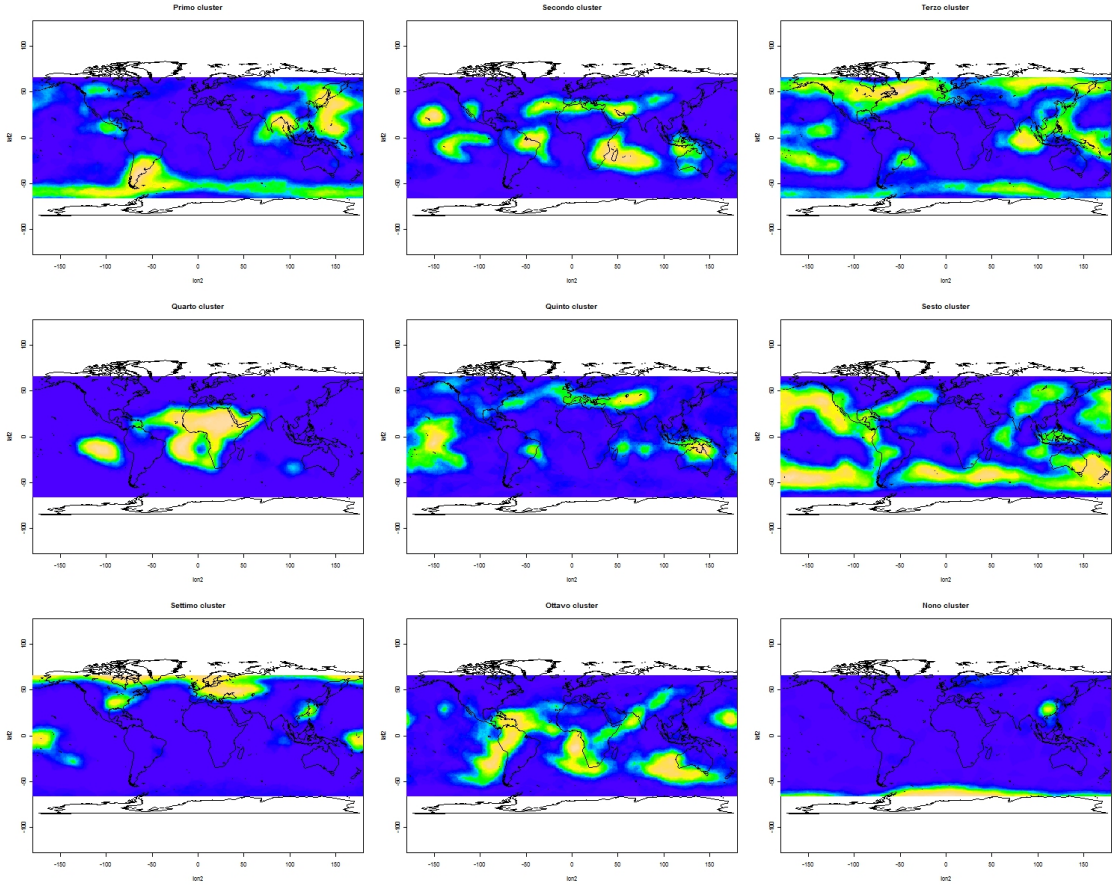


Figura 4.36: Frequenza di assegnazione dei siti della mappa ai diversi cluster lungo le iterazioni, per $k = 9$; i valori vicini a 1 sono mostrati in bianco, quelli vicino a 0 in blu

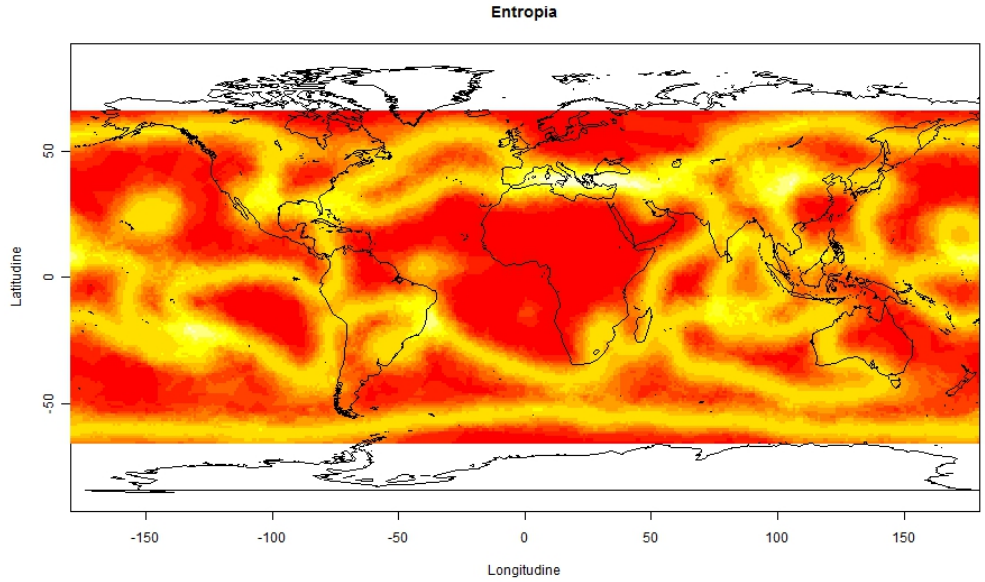


Figura 4.37: Entropia della classificazione ottenuta con il metodo dello spatial clustering per $k = 3$; i valori vicini a 1 sono mostrati in bianco, quelli vicini a 0 in rosso

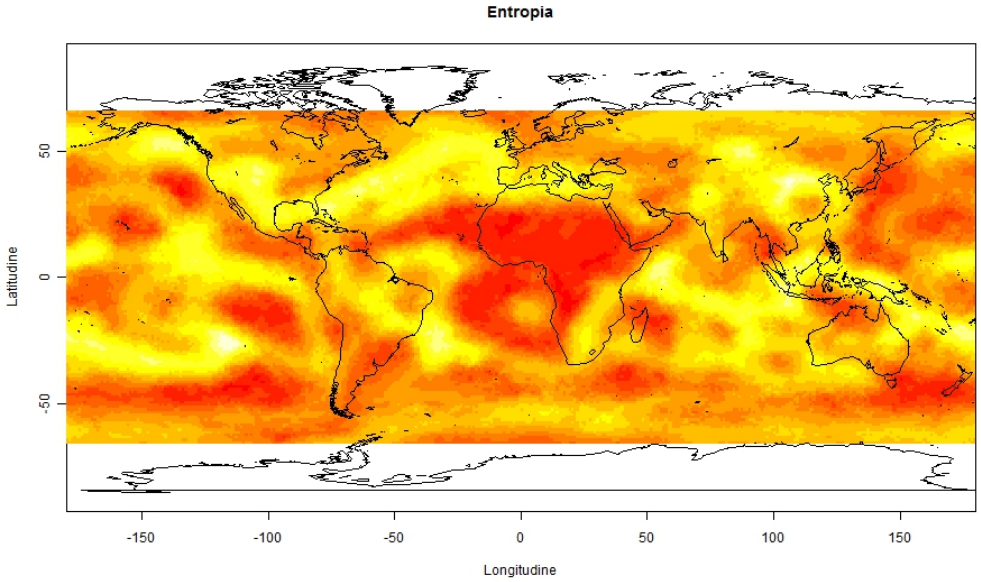


Figura 4.38: Entropia della classificazione ottenuta con il metodo dello spatial clustering per $k = 6$; i valori vicini a 1 sono mostrati in bianco, quelli vicini a 0 in rosso

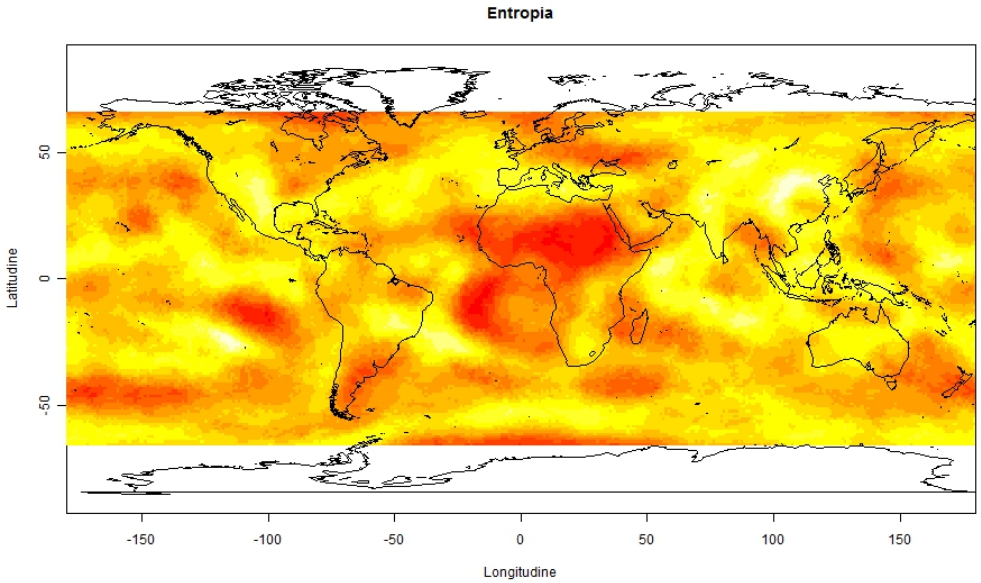


Figura 4.39: Entropia della classificazione ottenuta con il metodo dello spatial clustering per $k = 9$; i valori vicini a 1 sono mostrati in bianco, quelli vicini a 0 in rosso

Notiamo che l'entropia è maggiore rispetto al caso precedente in molte zone, anche interne ai cluster ottenuti dalla stima finale in Figura 4.32, anche perché i cluster ottenuti con $k = 6$ sono necessariamente meno estesi di quelli ottenuti per $k = 3$.

Osservando infine l'entropia ottenuta per $k = 9$ (Figura 4.39) vediamo che i confini tra i diversi cluster non si riescono a distinguere nella maggior parte delle zone della mappa, e in generale l'entropia è molto aumentata rispetto ai due casi precedenti. L'unica zona dove l'entropia è bassa è quella relativa al cluster che comprende il continente africano, (il cluster verde di Figura 4.33) che è effettivamente il cluster geograficamente più esteso[2].

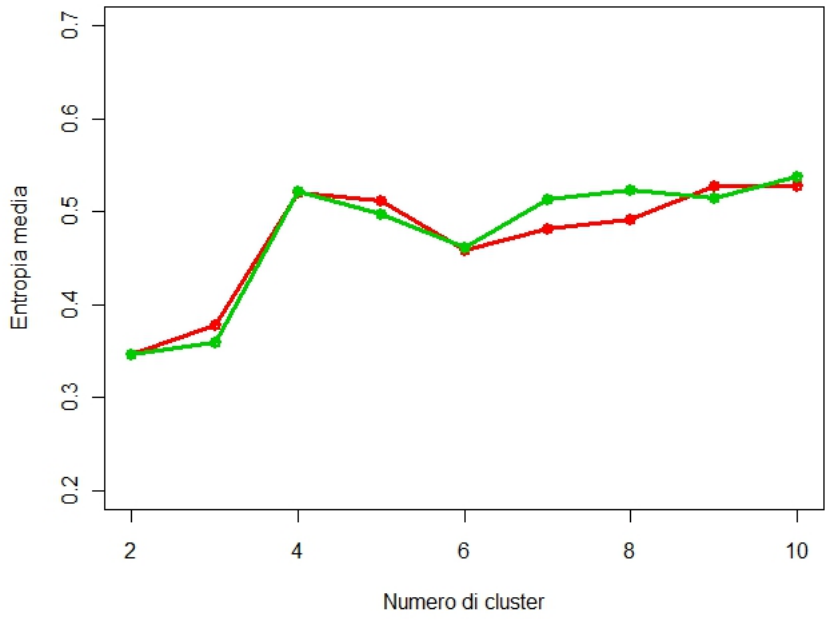


Figura 4.40: Andamento dell'entropia media all'aumentare del numero di cluster: in verde per classificazioni ottenute con $q = 3$ componenti principali, in rosso con $q = 5$

In Figura 4.40, infine, è riportato l'andamento dell'entropia media su tutta la mappa all'aumentare del numero di cluster, da $k = 2$ fino a $k = 10$. Le due curve rappresentano due diverse riduzioni dimensionali dei dati cha abbiamo utilizzato: in verde, la tecnica scelta, che utilizza $q = 3$ componenti principali per classificare i dati; in rosso, invece, mostriamo gli stessi risultati ottenuti mantenendo $q = 5$ componenti principali per classificare i dati.

Osservando le due curve, notiamo innanzi tutto che la quarta e la quinta componente principale danno un apporto molto scarso alla classificazione (che non abbiamo riportato per brevità). I valori dell'entropia media rilevati, come anche i risultati delle stime finali della classificazione, sono sostanzialmente gli stessi sia per $q = 3$ che per $q = 5$. La classificazione dei dataset composti dai dati di sintesi dei tasselli, quindi, viene effettuata esclusivamente grazie alle prime tre componenti principali.

[2]Si ricordi che, nel guardare le mappe mostrate, le aree sono deformate: quelle poste a latitudini elevate vengono ingrandite e quelle poste a latitudini equatoriali vengono rimpicciolite (come si era osservato nel Paragrafo 4.3.1)

Per quanto riguarda l'andamento del grafico, notiamo che, come avevamo immaginato, l'entropia ha un andamento generalmente crescente all'aumentare di k. Nel caso preso in esame, tuttavia, tale andamento non è monotono: l'entropia cresce leggermente tra $k = 2$ e $k = 3$, probabilmente a causa del fatto che con più cluster abbiamo più zone di confine elle quali l'entropia è maggiore. Invece, da 3 a 4 l'aumento del'entropia molto più ripido suggerisce che con 4 cluster l'entropia non aumenta esclusivamente a causa dell'incremento delle zone di confine, ma anche perché le classificazioni a diverse iterazioni sono più instabili. Infatti, aumentando ulteriormente il valore di k, l'entropia diminuisce, fino a trovare un minimo locale per $k = 6$. A partire da tale valore, l'andamento è di nuovo crescente.

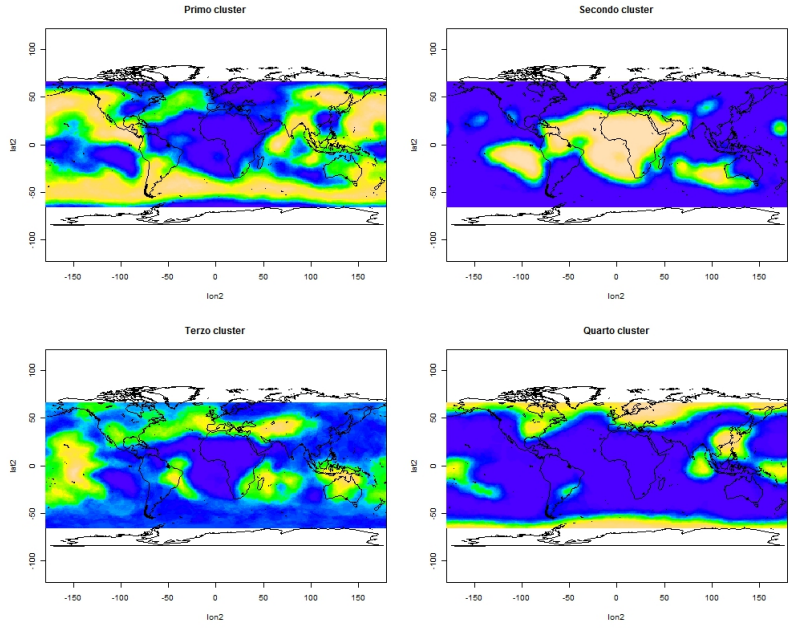


Figura 4.41: Frequenza di assegnazione dei cluster per $k = 4$; da bianco a blu: valori da 1 a 0

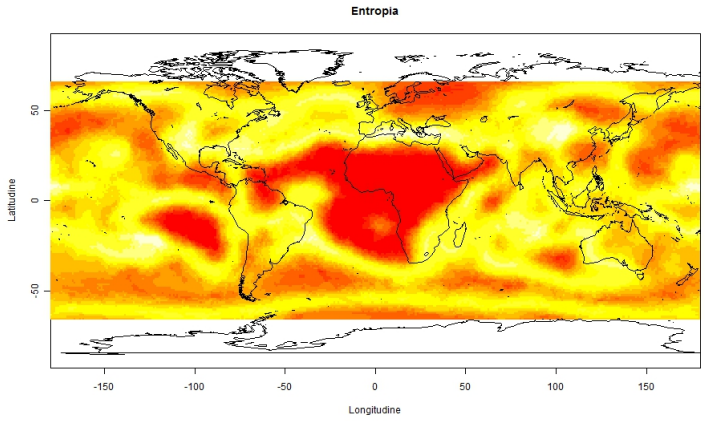


Figura 4.42: Entropia della classificazione ottenuta per $k = 4$; da bianco a rosso: valori da 1 a 0

La classificazione ottenuta per $k = 4$ risulta quindi essere un massimo locale per

l'entropia media. Osserviamo nelle Figure 4.41 e 4.42, rispettivamente le frequenze di assegnazione dei cluster e la mappa dell'entropia nel caso $k = 4$. Confrontando le mappe delle frequenze di assegnazione vediamo che, in effetti, per $k = 3$ o $k = 6$ i cluster trovati erano ben definiti, e non c'erano sulla mappa zone in cui le frequenze di assegnazione erano intermedie, fatta esclusione per le zone di confine tra i cluster. Nel caso $k = 4$, invece, ci sono due cluster ben definiti (il secondo ed il quarto), mentre gli altri due (soprattutto il terzo) presentano zone incerte molto al di fuori di quelli che sono i confini finali dei cluster.

Questo problema, ovviamente, si ripercuote pesantemente nella mappa dell'entropia della classificazione (Figura 4.42), dove si vede un'unica zona nella quale l'entropia è bassa, cioè la zona interna al secondo cluster. In tutte le altre zone della mappa, l'entropia è mediamente più alta di quella ottenuta per $k = 3$ e $k = 6$.

Dall'osservazione della Figura 4.42 scaturisce un'importante osservazione: il motivo per cui per $k = 4$ troviamo un massimo locale per l'entropia media, è la struttura probabilmente annidata dei dati stessi. Abbiamo osservato, infatti che per $k = 3$ la suddivisione in cluster viene effettuata esclusivamente secondo la prima componente principale. Per $k = 4$, invece, anche la seconda componente principale entra in gioco nella suddivisione in cluster dei dati di sintesi. Tuttavia, almeno due cluster del risultato ottenuto per $k = 3$ devono essere suddivisi nella direzione della seconda componente principale, ma per fare ciò 4 cluster non sono sufficienti, ce ne vogliono 5 o 6. Per questo motivo, la classificazione dei dati di sintesi ottenuta per $k = 4$ è molto instabile, e si osservano valori dell'entropia mediamente più elevati.

Le osservazioni espresse in questa sezione riguardo all'andamento dell'entropia suggeriscono che il numero di cluster che porta ad una migliore classificazione può essere scelto tra i due valori $k = 3$ e $k = 6$. Per $k = 3$ i risultati sono migliori, ma ciò può dipendere dal fatto che per $k = 6$ le aree di confine tra i cluster sono maggiori, e i cluster stessi sono meno estesi.

La scelta finale della classificazione ottima dipende anche dalle finalità dell'analisi effettuata, e dal grado di precisione richiesto. Se l'obiettivo è una suddivisione molto ampia e di facile interpretazione, il risultato ottenuto per $k = 3$ può essere considerato ottimale, poiché i gruppi risultanti si interpretano semplicemente come zone aride con poca copertura nuvolosa, temperate con copertura nuvolosa media e piovose o con copertura nuvolosa maggiore. Se, al contrario, si cerca una suddivisione più precisa, il risultato migliore è quello per $k = 6$, nel quale si tiene conto anche della stagionalità per effettuare la suddivisione.

4.4 Classificazione dei dati con un metodo gerarchico

Nel Capitolo 2 abbiamo descritto una procedura di classificazione per dati funzionali spazialmente dipendenti, proponendo per la classificazione l'utilizzo del metodo delle k–medie. Nulla vieta, tuttavia, di classificare i vettori degli scores utilizzando altri metodi di classificazione. Inoltre, abbiamo più volte osservato, nel corso della Sezione 4.3, che la struttura finale delle mappe di classificazione ottenute sembra suggerire una disposizione annidata dei cluster. Nel corso di questa sezione, vogliamo quindi esplorare in maniera più approfondita tale supposizione, per verificare se effettivamente esiste tale struttura, tramite classificazione con una procedura gerarchica.

Ripetiamo quindi le stesse analisi della sezione precedente, utilizzando ad ogni iterazione una tecnica di clustering gerarchica per la classificazione dei vettori degli scores associati ai dati di sintesi dei tasselli (per un'analisi approfondita dei metodi gerarchici

di classificazione multivariata, si veda [Johnson and Wichern, 2002]). La scelta dell'utilizzo di un metodo gerarchico per il clustering dei dati assicura che, ad ogni iterazione, la struttura dei cluster per diversi valori di k sia annidata: i cluster ottenuti per $k = 3$ sono ottenuti dal risultato di $k = 2$, dividendo in due uno dei due cluster esistenti, e così via.

Osserviamo che, tuttavia, tale metodo non garantisce che le stime finali ottenute siano, effettivamente, annidate: ricordiamo, infatti, che per ottenere la stima finale scegliamo, per ogni pixel, il cluster al quale è stato associato più volte nel corso delle diverse iterazioni del metodo di classificazione. In questo modo, anche se le clusterizzazioni ottenute per le diverse iterazioni, prese singolarmente, hanno una struttura annidata, tale proprietà non è garantita per la stima finale. Applichiamo, comunque, tale metodo di classificazione ai dati, per verificare le differenze nella classificazione finale ottenuta rispetto ai risultati che abbiamo presentato nel caso delle k–medie, nella Sezione 4.3.

Coerentemente a quanto mostrato per la classificazione tramite k–medie, applichiamo, ad ogni iterazione, la classificazione con metodo gerarchico agli scores relativi alle prime tre componenti principali dei dati di sintesi dei tasselli. Utilizziamo come distanza la semimetrica in L^2 indotta dalle prime $q = 3$ autofunzioni, che equivale alla distanza euclidea tra i vettori degli scores; per quanto riguarda la scelta del metodo di linkage, utilizziamo il metodo *Ward*[3] che è quello che permette di ottenere, nel caso dei dati in esame, i risultati migliori.

Come nella sezione precedente, analizziamo in un primo momento i risultati ottenuti in una singola iterazione del metodo, per poi commentare i risultati finali.

Analisi di una singola iterazione

Mostriamo ora i risultati della classificazione, con il metodo gerarchico specificato, dei dati di sintesi relativi ad una particolare tassellazione di Voronoi.

In Figura 4.43 osserviamo innanzi tutto il dendrogramma ottenuto, che fornisce una rappresentazione grafica del processo gerarchico di assegnazione.

Osserviamo che la classificazione dei dati con tale metodo è molto chiara: si ottiene infatti un buon dendrogramma, nel quale si riesce visivamente ad avere un'idea della struttura della clusterizzazione. Dalla Figura, poi, vediamo che la clusterizzazione più netta, e quindi migliore, sembra essere quella ottenuta per $k = 3$. Ricordiamo, infatti, che in ordinata di un dendrogramma è rappresentato il livello di distanza tra i diversi cluster in corrispondenza del quale tali cluster vengono raggruppati. Si osserva quindi una classificazione migliore per k cluster se, nel dendrogramma, si osserva che per un'ampio intervallo di valori in ordinata si trovano esattamente k rami.

Nelle Figure 4.44 e 4.45, mostriamo quindi la classificazione degli scores effettuata con clustering gerarchico, rispettivamente per $k = 3$ e $k = 6$. Dalle Figure notiamo che, effettivamente, i cluster hanno struttura annidata, come ci aspettavamo. Inoltre, diversamente da quanto ottenuto per le k–medie, notiamo che già da $k = 3$ la clusterizzazione viene effettuata tenendo conto delle prime due componenti principali, mentre la terza non sembra essere mai utilizzata.

Un'altra osservazione che possiamo fare, confrontando le Figure 4.44 e 4.45 con quanto avevamo ottenuto con il metodo delle k–medie (Figure 4.24 e 4.25), nella clas-

[3]Il linkage Ward utilizza come criterio per agglomerare due diversi cluster la decomposizione della varianza totale in varianza within e between: si parte da $n = 300$ cluster distinti, e ad ogni passo viene fusa la coppia di cluster che porta a un minore incremento della varianza within

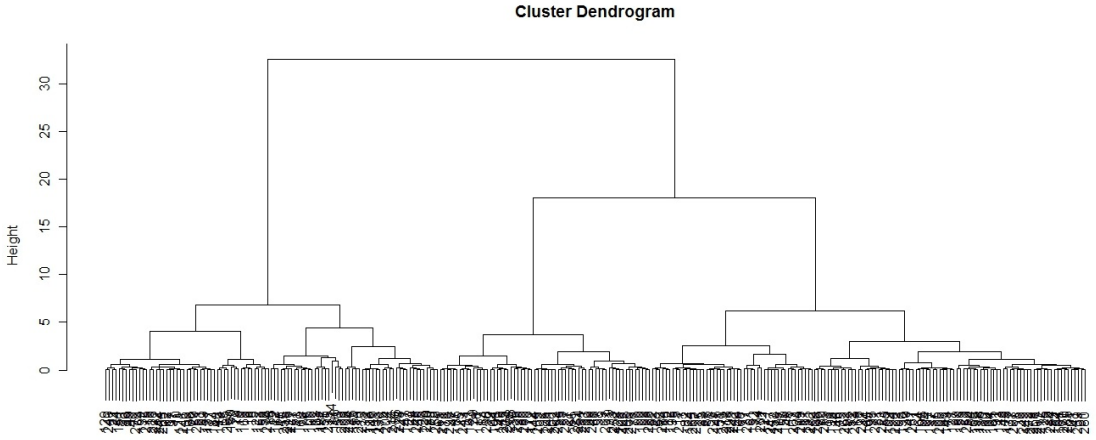


Figura 4.43: Dendrogramma relativo alla classificazione gerarchica degli scores delle prime tre componenti principali dei dati di sintesi di una tassellazione di Voronoi, con distanza euclidea, e linkage Ward

sificazione si tiene conto dell'interazione tra le prime due componenti, e il risultato ottenuto sembra visivamente migliore.

Risultati dopo $M = 100$ iterazioni

Osserviamo, infine, i risultati ottenuti dopo $M = 100$ iterazioni del metodo proposto. Nelle Figure 4.46 e 4.47 mostriamo il confronto delle stime finali ottenute, per $k \in \{2, 3, ..., 7\}$, ovvero le classificazioni finali ottenute con k–medie (pannelli di sinistra) e clustering gerarchico con distanza euclidea e linkage Ward (pannelli di destra).

Osserviamo che, in tutti i casi, le stime che abbiamo ottenuto sono molto robuste al variare della tecnica di classificazione utilizzata; questo fatto conferma la validità dei risultati ottenuti, che non sembrano quindi dipendere dal particolare metodo scelto per la classificazione.

L'unico caso in cui la classificazione finale ottenuta presenta delle sostanziali differenze è il caso $k = 4$ (ultima riga in Figura 4.46). In questo caso, infatti, nella classificazione finale ottenuta con clustering gerarchico uno dei cluster viene assegnato a pochissimi siti. Osservando le frequenze di assegnazione ai cluster per quel caso (mostrate in Figura 4.48) notiamo che questo succede perché il quarto cluster viene associato, nel corso delle diverse iterazioni del metodo, a zone sempre diverse, e nel risultato finale non compare, se non in piccole zone molto dubbie. Tale osservazione poteva essere fatta anche a partire dal dendrogramma ottenuto in Figura 4.43. Osservando il dendrogramma dall'alto in basso notiamo infatti che la suddivisione per $k = 5$ segue a poca distanza quella ottenuta per $k = 4$. In questo caso, quindi, il metodo, suddivide in modi diversi in ogni iterazione i tre cluster ottenuti per $k = 3$, causando un aumento del'entropia. A questo proposito, ricordiamo che per la classificazione effettuata tramite k–medie, avevamo osservato che la classificazione per $k = 4$ era molto instabile (massimo dell'entropia), aspetto che ritroviamo anche nel caso della classificazione gerarchica, e che risulta potenziato dal fatto che ora i cluster sono vincolati ad essere annidati ad ogni iterazione.

Infine, osserviamo in Figura 4.49 l'andamento dell'entropia media all'aumentare del numero di cluster con k–medie (curva verde) e clustering gerarchico (curva blu).

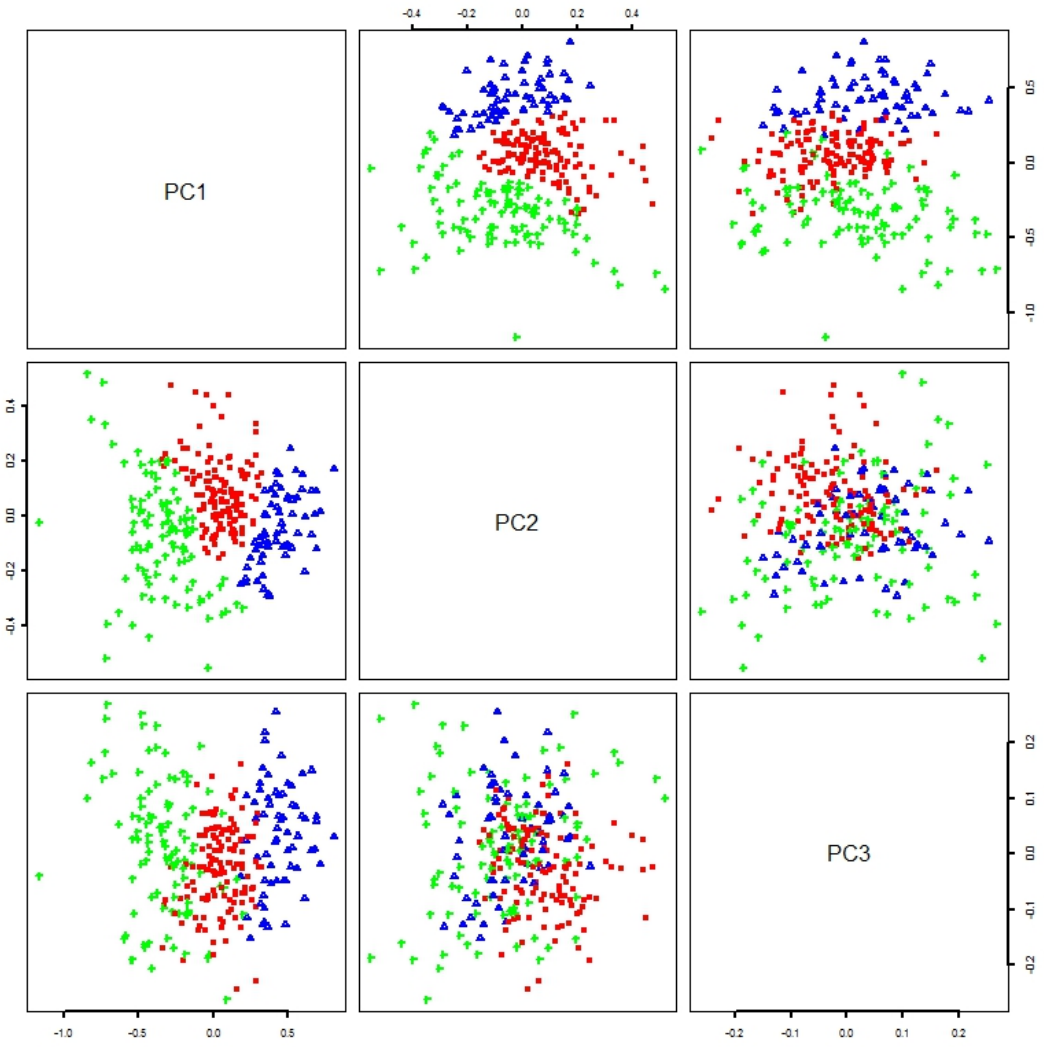


Figura 4.44: Scatterplot degli scores delle prime tre componenti principali, colorati a seconda della classificazione ottenuta per $k = 3$ con metodo gerarchico

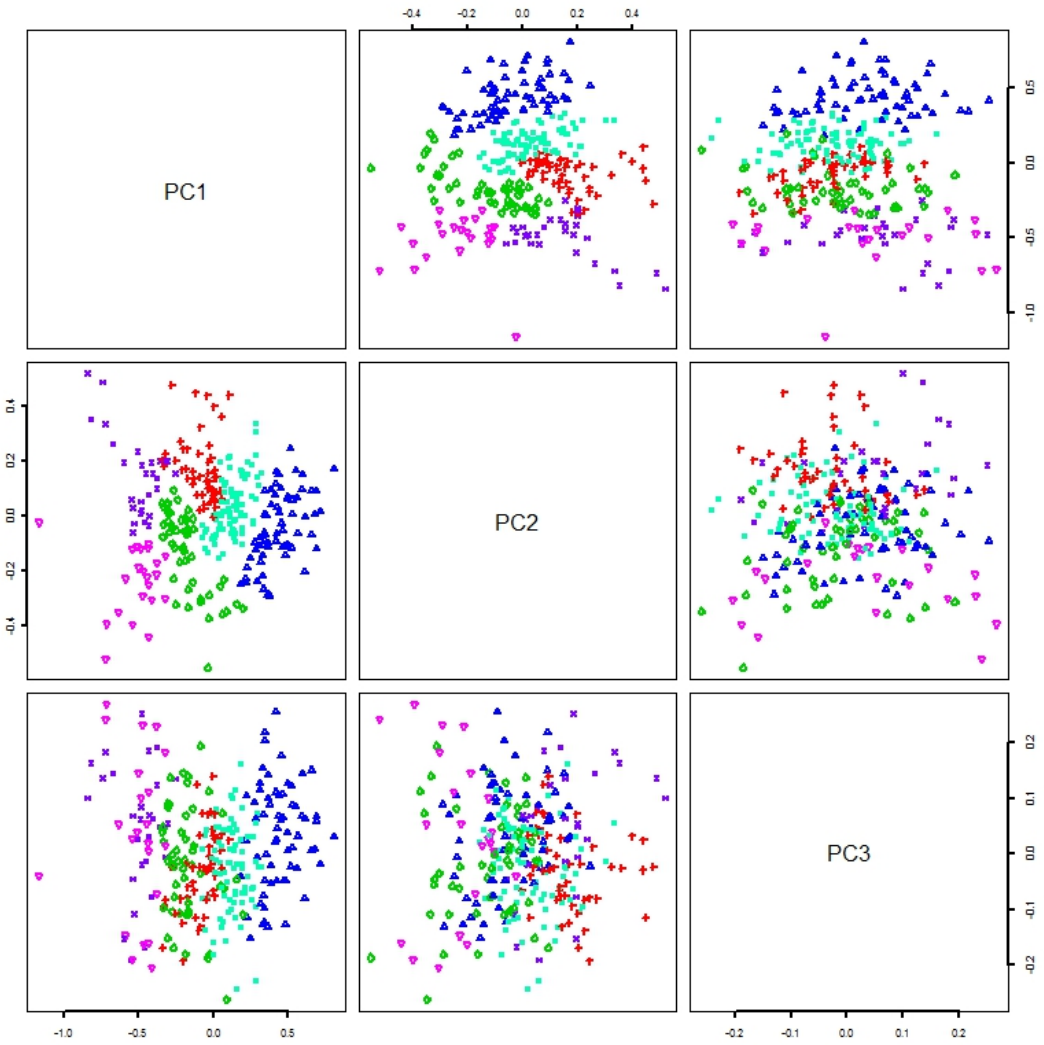


Figura 4.45: Scatterplot degli scores delle prime tre componenti principali, colorati a seconda della classificazione ottenuta per $k = 6$ con metodo gerarchico

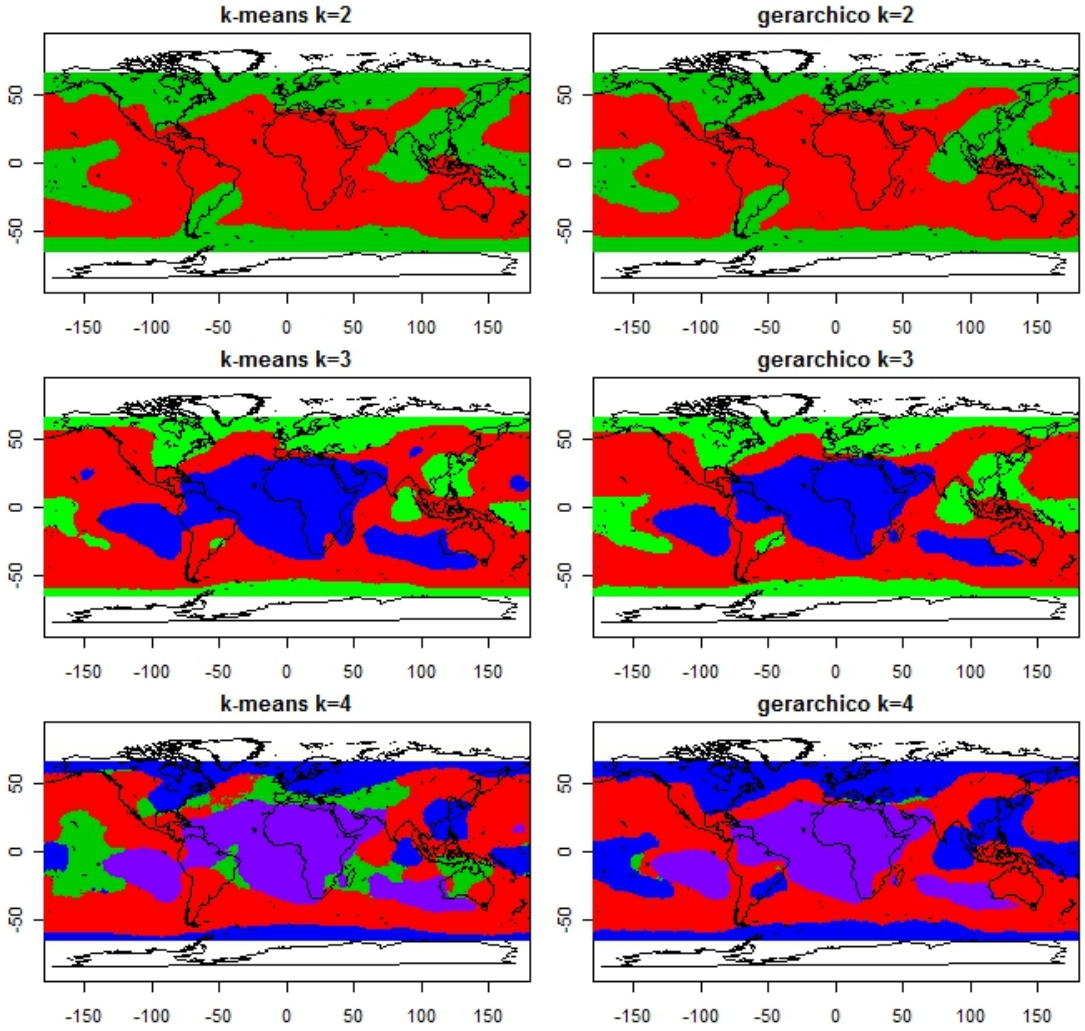


Figura 4.46: Confronto della mappa finale di clusterizzazione ottenuta con k–medie (pannelli di sx) e clustering gerarchico (pannelli di dx), per $k = 2, 3, 4$, sugli scores della FPCA

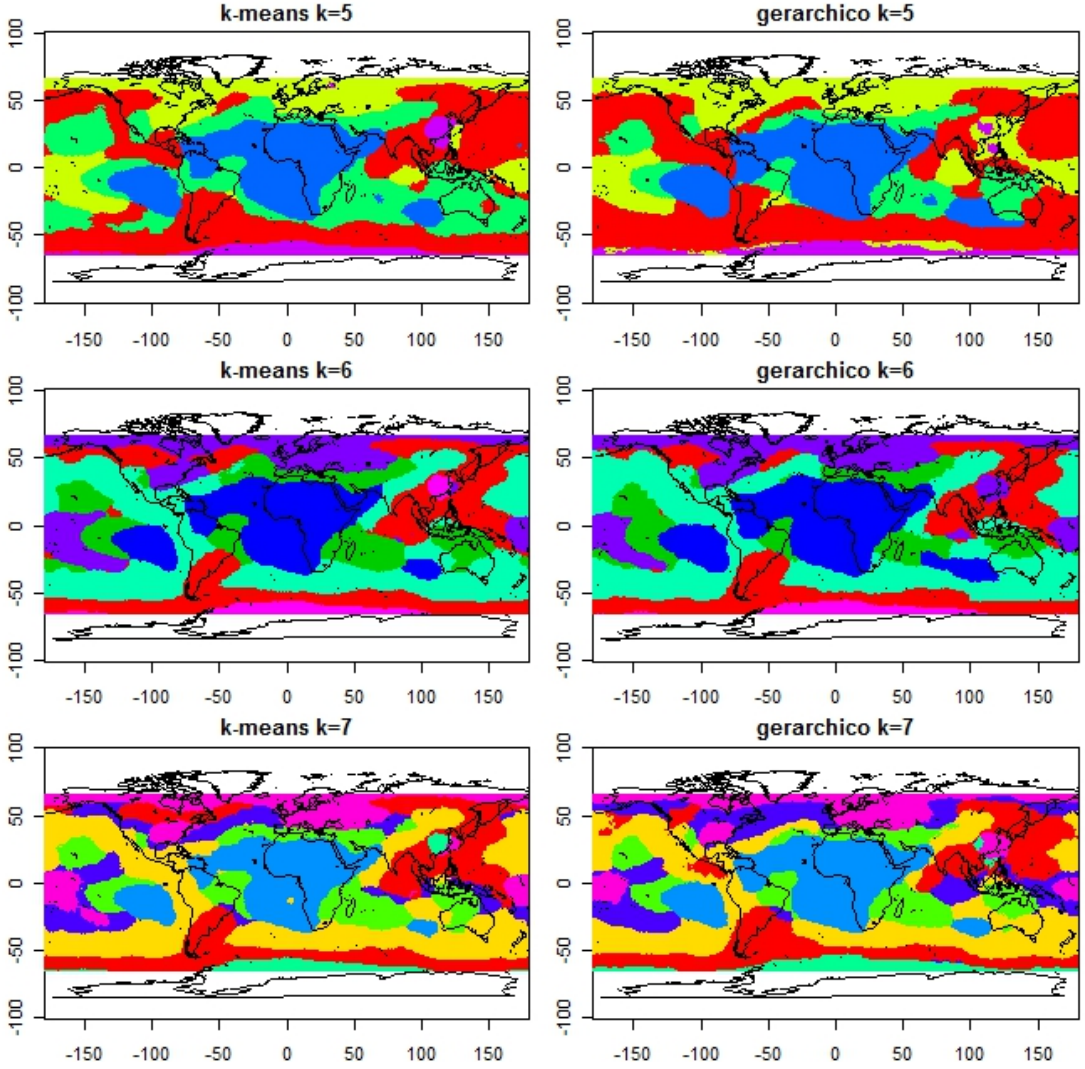


Figura 4.47: Confronto della mappa finale di clusterizzazione ottenuta con k–medie (pannelli di sx) e clustering gerarchico (pannelli di dx), per $k = 5, 6, 7$, sugli scores della FPCA

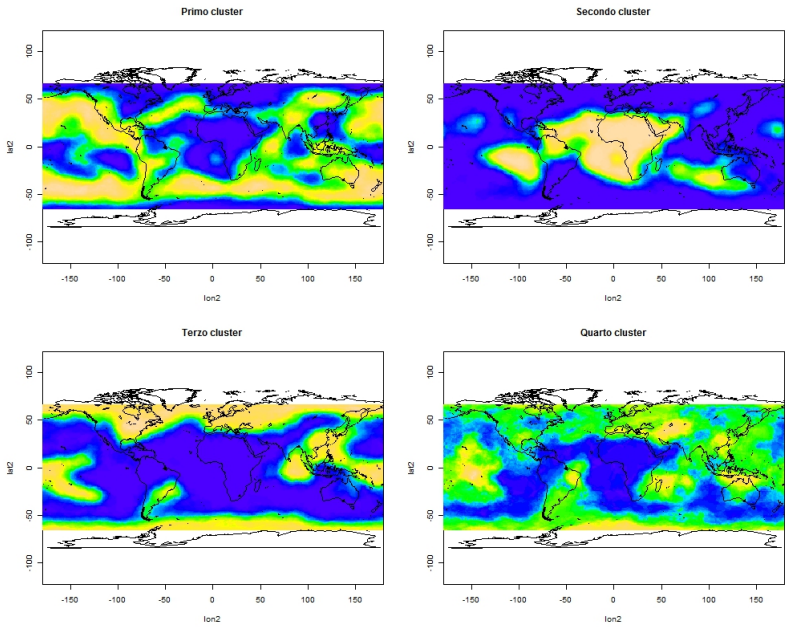


Figura 4.48: Frequenza di assegnazione dei siti della mappa ai diversi cluster lungo le iterazioni, per $k = 4$ con metodo gerarchico; i valori vicini a 1 sono mostrati in bianco, quelli vicino a 0 in blu

Osserviamo innanzitutto che l'entropia delle classificazioni ottenute tramite clustering gerarchico è più elevata rispetto a quanto ottenuto tramite k–medie. Questo vuol dire, quindi, che le classificazioni delle diverse iterazioni del metodo proposto effettuate con clustering gerarchico sono più instabili. Tale aspetto non deve sorprendere. Basta infatti osservare che, mentre utilizzando le k–medie ogni classificazione, al variare del numero di cluster, è indipendente dall'altra, per il clustering gerarchico la classificazione ottenuta con un certo $\bar{k}$ condiziona fortemente quelle ottenute per $k > \bar{k}$, che sono vincolate ad essere annidate. Di conseguenza, se si genera una tassellazione di Voronoi per la quale la classificazione dei dati di sintesi è molto diversa rispetto alle precedenti per un valore basso di k, tale differenza si riflette necessariamente sui risultati relativi alla stessa iterazione per valori maggiori di k, aumentando l'instabilità tra diverse iterazioni.

Il risultato interessante è che, sebbene i valori di entropia ottenuti siano maggiori utilizzando il clustering gerarchico, l'andamento dell'entropia all'aumentare del numero di cluster è lo stesso nei due casi. L'unica differenza che osserviamo è che, ora, non troviamo più un minimo per $k = 6$, ma un andamento abbastanza costante da $k = 4$ in poi.

Alla luce degli ultimi risultati, possiamo affermare che le classificazioni ottenute sono molto robuste al variare delle tecniche di classificazione utilizzate, aspetto che conferma la validità delle stime finali ottenute. Tali stime, infatti, non dipendono dal particolare metodo scelto per la classificazione dei dati di sintesi: non si tratta, quindi, di classificazioni fittizie create dal particolare algoritmo utilizzato, ma al contrario, i cluster ottenuti descrivono il reale comportamento dei dati.

Anche con l'utilizzo di una tecnica di classificazione gerarchica, non osserviamo esattamente una struttura annidata per le clusterizzazioni finali, che tuttavia continuano a mantenere delle forme e strutture riconoscibili al variare del numero di cluster utilizzato.

Infine, osserviamo che la classificazione migliore ottenuta è certamente quella con

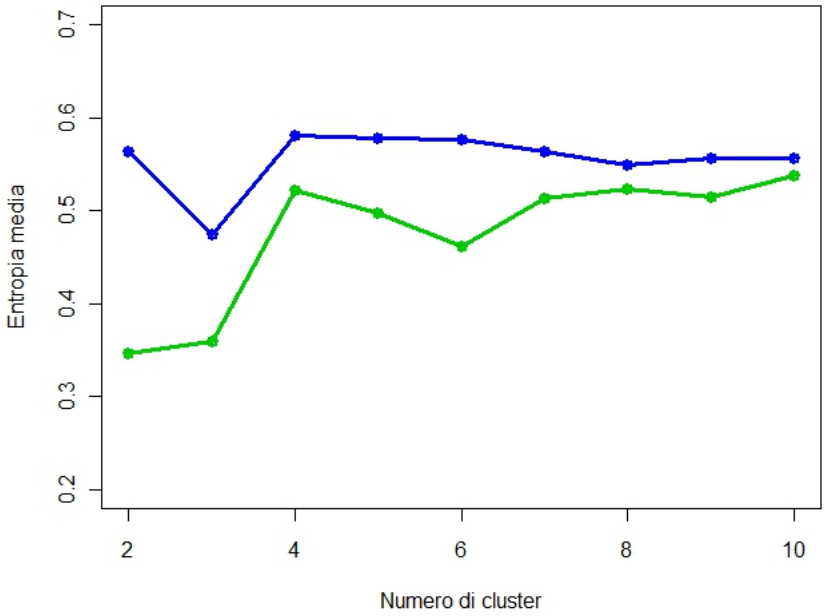


Figura 4.49: Andamento dell'entropia media all'aumentare del numero di cluster: in verde per classificazioni ottenute con k–medie, in blu con clustering gerarchico

soli tre cluster, e che anche tale risultato è robusto al variare delle tecniche utilizzate. Per quanto riguarda, invece, l'altro minimo di entropia che avevamo ottenuto per $k = 6$, notiamo che tale risultato è stato ottenuto esclusivamente utilizzando il metodo delle k–medie per classificare i dati di sintesi. Utilizzando invece una tecnica gerarchica, non osserviamo un minimo per $k = 6$, anzi, l'entropia diminuisce leggermente passando a 7 o 8 cluster.

4.5 Analisi dell'energia massima richiesta dal sistema di backup

4.5.1 Un'altra variabile di interesse

Nelle sezioni precedenti, abbiamo analizzato il database NASA Earth Surface Meteorology and Solar Energy, con un'attenzione particolare verso un importante parametro in esso contenuto, e cioè il numero equivalente di giorni consecutivi senza sole. La classificazione dei dati relativi a tale parametro ha fornito una suddivisione della superficie terrestre in aree nelle quali l'evoluzione annuale di tale parametro è omogenea. In particolare, la classificazione ottenuta per $k = 3$ cluster, suggerisce una divisione in tre aree fondamentali: una prima regione, nella quale i periodi di copertura nuvolosa sono mediamente più lunghi, una seconda regione nella quale i periodi di copertura nuvolosa hanno lunghezza intermedia, ed una terza regione nella quale tali periodi sono mediamente più brevi, durante tutto l'anno.

Le suddivisioni ottenute si basano esclusivamente sulla lunghezza dei periodi senza sole. Non si tiene conto quindi, per esempio, dell'insolazione media mensile registrata

nelle diverse zone. Tali suddivisioni, quindi, sono rilevanti dal punto di vista climatico, ed è importante considerare i risultati ottenuti per la scelta della regione più adatta per il posizionamento di un impianto fotovoltaico; per la progettazione e il dimensionamento di tali impianti, tuttavia, è necessario prendere in considerazione anche altri parametri. Per lo stesso problema di dimensionamento delle batterie ausiliarie collegate agli impianti fotovoltaici che è stato posto all'inizio del capitolo, per esempio, è importante considerare un altro parametro derivato dall'insolazione, e cioè l'energia massima richiesta, mensilmente, dal sistema di backup. Osserviamo infatti che, sebbene il numero equivalente di giorni senza sole sia un parametro di riferimento, non contiene informazioni sulla quantità massima di energia che è necessario immagazzinare, o fornire attraverso un sistema esterno, ma solo sui periodi massimi di funzionamento di tale sistema.

Per ottenere, quindi, la quantità massima di energia che le batterie ausiliarie devono poter immagazzinare, è necessario moltiplicare il numero equivalente di giorni consecutivi senza sole per l'insolazione media mensile, per ogni sito. Così facendo, otteniamo un dataset contenente le informazioni cercate, sullo stesso reticolo regolare considerato precedentemente.

In questa sezione ci concentreremo, quindi, sull'analisi e in particolare sulla classificazione, utilizzando la procedura di spatial clustering precedentemente analizzata, dei dati derivanti dal calcolo della quantità di interesse a partire dalle informazioni a nostra disposizione. Effettuando la classificazione spaziale di questo dataset, otterremo come risultato finale un'interpretazione delle aree associate allo stesso cluster in termini di zone che richiedono uno stesso tipo di batteria esterna da collegare agli impianti.

Come per l'analisi precedente, anche in questo caso è necessario escludere le calotte polari dall'analisi. Infatti, i dati che ci proponiamo di analizzare in questa sezione sono ottenuti semplicemente moltiplicando, sito per sito, i dati relativi all'insolazione mensile (disponibili su tutto il reticolo) per i dati relativi al numero equivalente di giorni senza sole consecutivi, che non è definito in corrispondenza dei siti associati alle calotte polari, come già spiegato nel Paragrafo 4.1.2.

Per rappresentare il dataset così ottenuto, scegliamo di mostrare nelle Figure 4.50 e 4.51 delle mappe rappresentanti la quantità di energia massima richiesta dal sistema di backup per ogni mese, nelle quali la quantità di energia relativa ad ogni sito è indicata dal colore del pixel ad esso associato. In particolare, utilizziamo una scala che va dal rosso, che indica siti in corrispondenza dei quali l'energia massima necessaria è quasi nulla, al blu, che indica siti in corrispondenza dei quali tale energia è vicina al valore massimo osservato.

Dalle 12 mappe osserviamo che i dati seguono una particolare struttura geografica: si intravedono, infatti, delle zone nelle quali i valori mensili sono simili. Tuttavia, l'alta variabilità del dataset rende necessario l'utilizzo di tecniche di classificazione che tengano conto anche delle informazioni spaziali al fine di ottenere una nitida e sensata identificazione di regioni omogenee. Tramite l'analisi dei dati funzionali associati alle osservazioni mensili della grandezza di interesse, vogliamo inoltre cogliere con un'unica etichetta per sito l'intera dinamica annuale del fenomeno, che è impossibile cogliere osservando le singole mappe relative ad ogni mese.

4.5.2 Scelte effettuate per lo smoothing e la classificazione dei dati

Nelle sezioni precedenti, abbiamo spiegato in dettaglio le scelte effettuate per lo smoothing e la classificazione dei dati. Vediamo quindi ora, brevemente, quali metodi sono stati scelti per l'analisi di questo nuovo dataset.

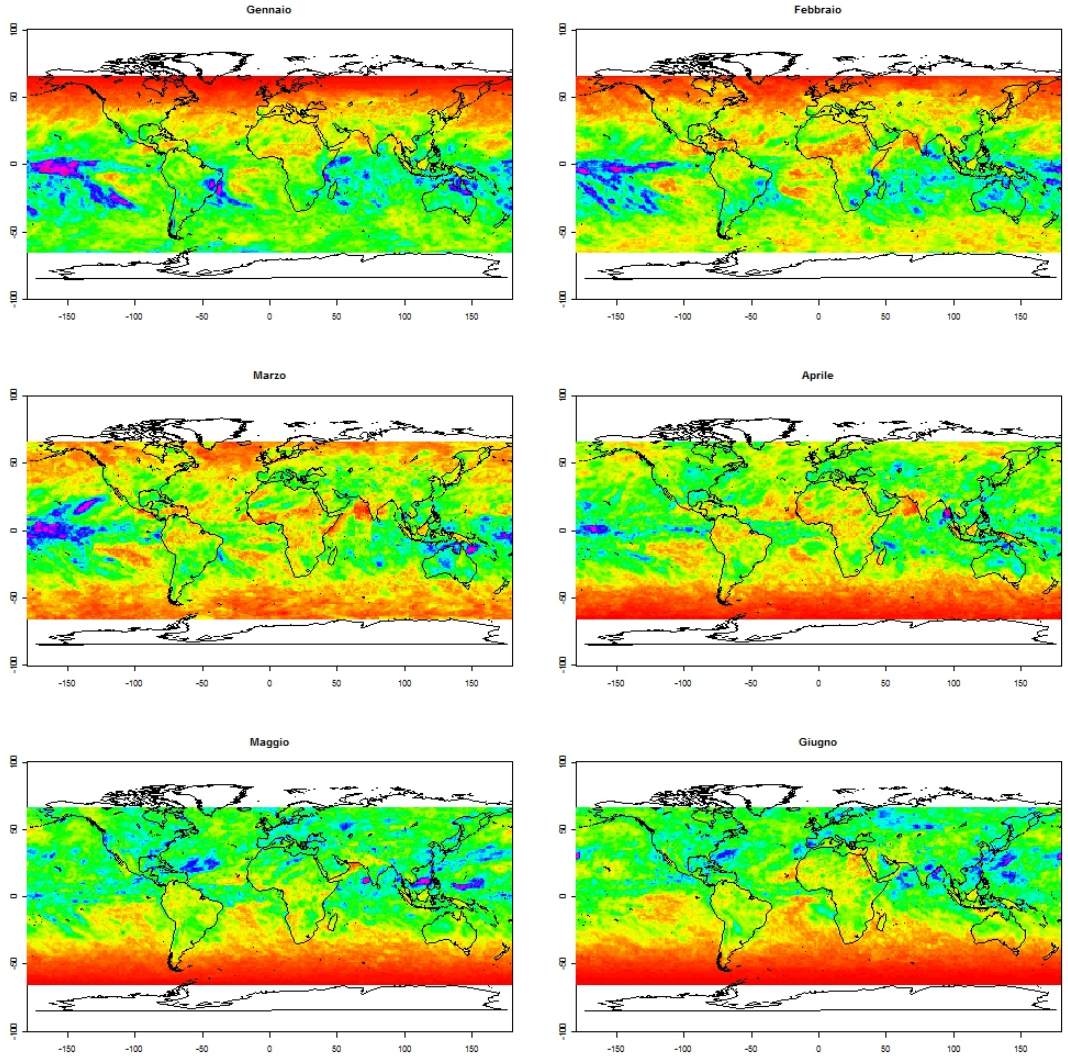


Figura 4.50: Mappe rappresentanti la quantità mensile di energia massima osservata negli ultimi 22 anni richiesta dal sistema di backup per ogni sito, nei mesi da gennaio a giugno; in rosso vengono rappresentati valori vicini allo zero, in blu valori vicini al valore massimo osservato (circa $92.5 kwh/m^2$)

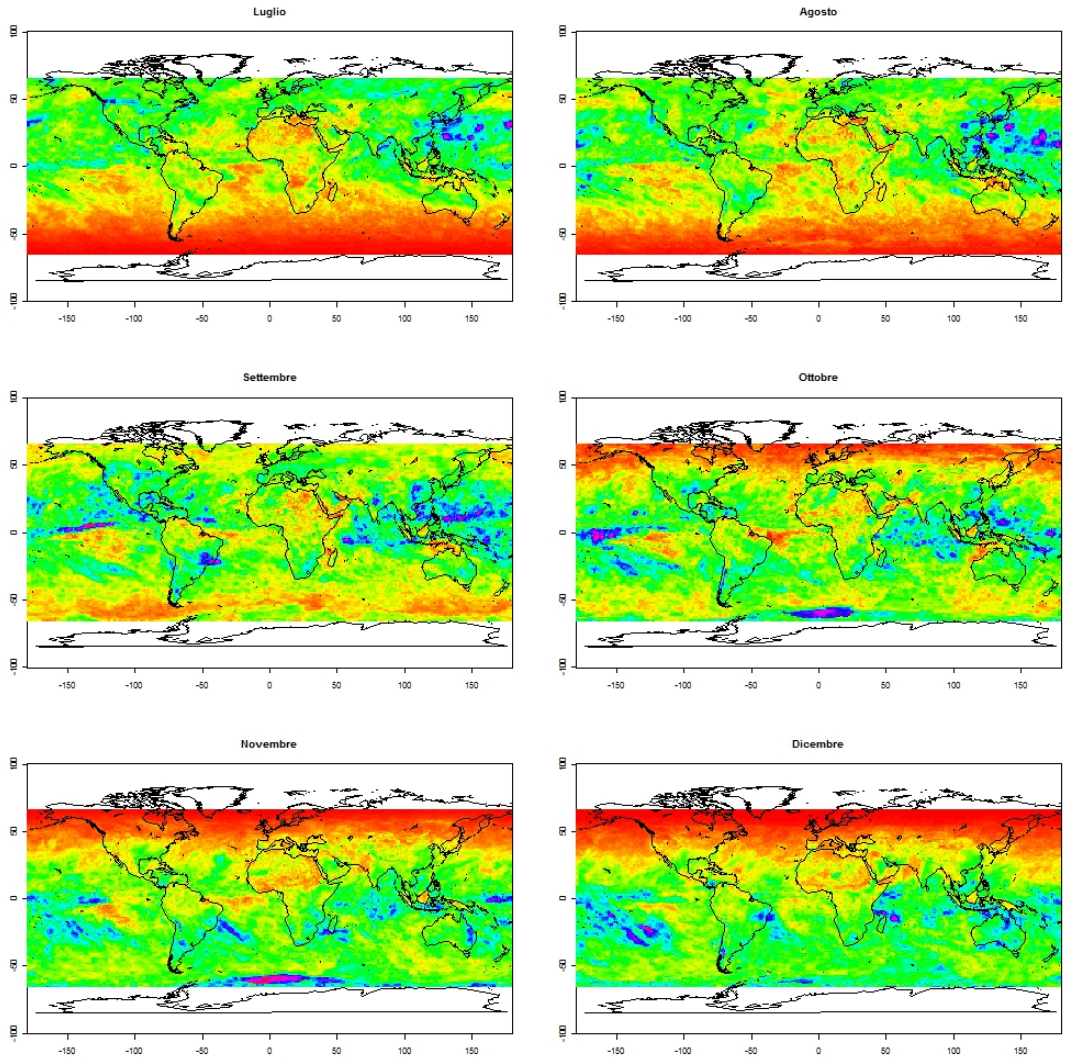


Figura 4.51: Mappe rappresentanti la quantità mensile di energia massima osservata negli ultimi 22 anni richiesta dal sistema di backup per ogni sito, nei mesi da luglio a dicembre; in rosso vengono rappresentati valori vicini allo zero, in blu valori vicini al valore massimo osservato (circa $92.5 kwh/m^2$)

Innanzitutto, al fine di riproporre la stessa analisi effettuata precedentemente sulla proporzione di giorni senza sole sul nuovo parametro di riferimento, è necessario applicare una tecnica di smoothing per passare dalle osservazioni discrete ai dati funzionali. Osserviamo che, anche in questo caso, il dato funzionale che ci proponiamo di ottenere è definito sul dominio $[0, 12]$, e deve avere un andamento periodico (o approssimativamente periodico) di periodo 12 mesi. Un'altra proprietà importante che richiediamo rispettino i dati è la positività: il dato funzionale rappresenta l'andamento annuale dell'energia massima immagazzinabile dal sistema di backup, quindi non sono ammessi dati negativi. A differenza dei dati precedentemente analizzati, invece, non è necessario richiedere che il dato sia minore di 1 in ogni punto del dominio, in quanto la grandezza analizzata è una quantità di energia assoluta, e non una proporzione.

Per effettuare lo smoothing di tali dati, in maniera analoga a quanto abbiamo proposto nella Sezione 4.2, utilizziamo un kernel gaussiano che garantisce che la proprietà di positività venga rispettata. Per quanto riguarda la scelta del bandwidth, osserviamo dai risultati ottenuti al variare di tale parametro in Figura 4.52 che anche in questo caso $h = 1.5$ permette una migliore rappresentazione dei dati funzionali.

Una volta creato il dataset di dati funzionali, per applicare l'algoritmo descritto nel Capitolo 2, è necessario scegliere una distanza tra i siti e un modo per effettuare la tassellazione di Voronoi, un metodo di riduzione dimensionale dei dati e un metodo di classificazione.

Per quanto riguarda la distanza tra i siti e la tassellazione di Voronoi, applichiamo lo stesso metodo proposto nel Paragrafo 4.3.1. I dati contenuti nel dataset che ci proponiamo di analizzare, infatti, sono osservazioni funzionali in corrispondenza dello stesso reticolo formato da meridiani e paralleli analizzato precedentemente. Scegliamo, quindi, di estrarre uniformemente dei centri sulla superficie terrestre, e generare a partire da tali centri una tassellazione di Voronoi utilizzando la distanza geodetica. Una volta ottenuta la tassellazione di Voronoi, procediamo come nell'analisi precedentemente effettuata, calcolando un dato di sintesi per ogni tassello.

Effettuiamo, poi, la riduzione dimensionale dei dati di sintesi, utilizzando il metodo della FPCA. Le autofunzioni ottenute in una singola iterazione dell'algoritmo sono presentate in Figura 4.53. Le prime due autofunzioni trovate spiegano, da sole più del 90% della variabilità totale, mentre le prime tre arrivano a spiegarne oltre il 95%. Osserviamo, poi, che le prime due autofunzioni ottenute identificano uno spazio formato da funzioni periodiche, di periodo uguale a 12 mesi, con un andamento approssimativamente sinusoidale. Con l'aggiunta della terza autofunzione, si aggiunge un'ulteriore componente periodica sfasata rispetto alla precedente.

Per quanto riguarda l'interpretazione delle componenti principali ottenute, osserviamo che la prima effettua un contrasto tra il periodo estivo e il periodo invernale, la seconda rappresenta semplicemente una modifica nella media dei dati, mentre la terza contrasta il periodo autunnale con quello primaverile. Le componenti principali successive alla terza presentano la stessa analogia con le funzioni della base di Fourier osservata nell'analisi della proporzione dei giorni consecutivi senza sole: il periodo delle autofunzioni ottenute si dimezza progressivamente. Osserviamo, tuttavia, che l'interpretazione di tali componenti è più difficile, e la variabilità spiegata decresce solo leggermente. Di conseguenza, scegliamo di procedere mantenendo le prime tre componenti principali per rappresentare i dati. Passiamo quindi da dati funzionali in uno spazio infinito dimensionale, alle proiezioni di tali dati su di uno spazio tridimensionale la cui base è formata dalle prime tre autofunzioni ottenute tramite la FPCA.

Per effettuare la classificazione dei dati di sintesi scegliamo di classificare le proiezioni

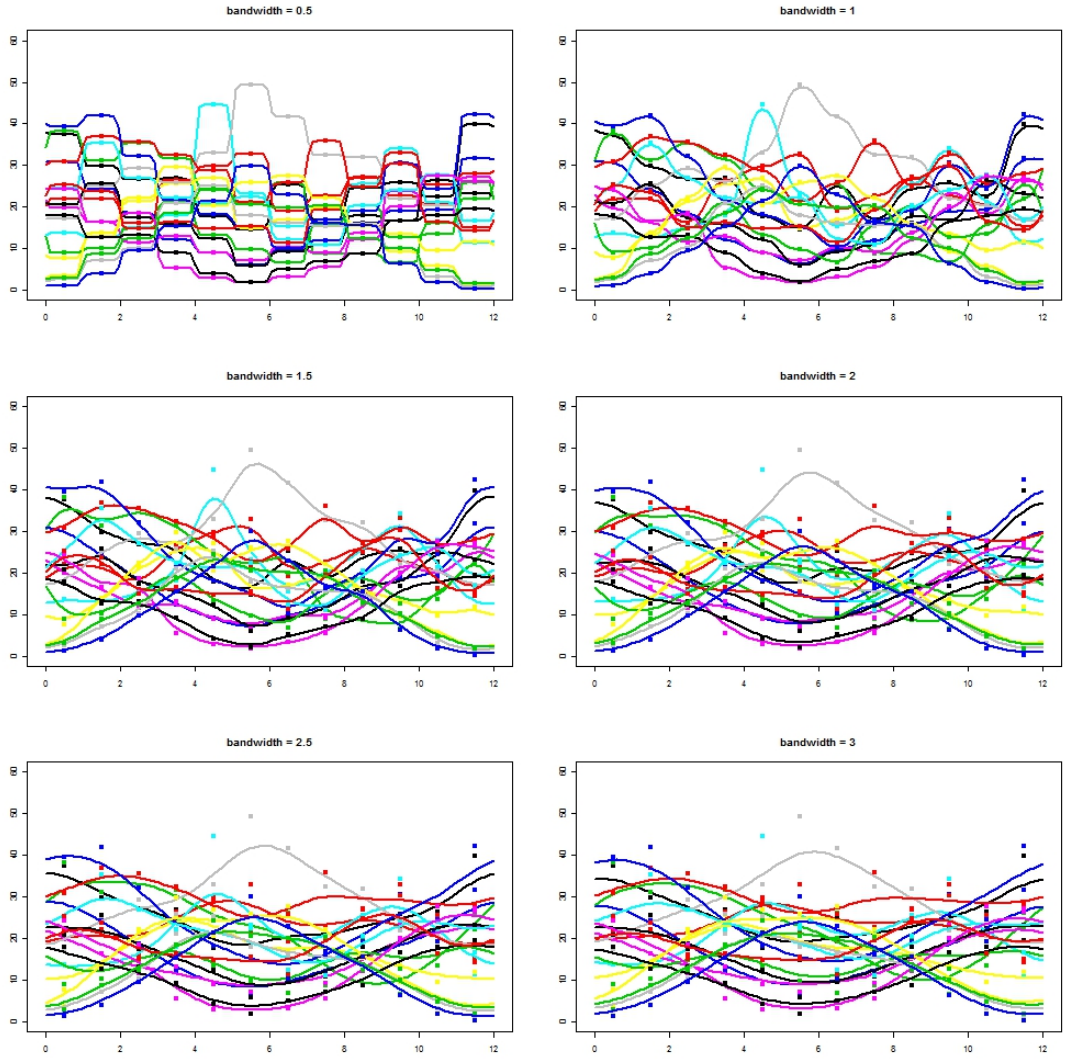


Figura 4.52: Dati funzionali ottenuti tramite smoothing con kernel gaussiano per 18 siti estratti casualmente dal dataset, da sinistra a destra e poi dall'alto in basso con bandwith $h = 0.5, 1, ..., 3$

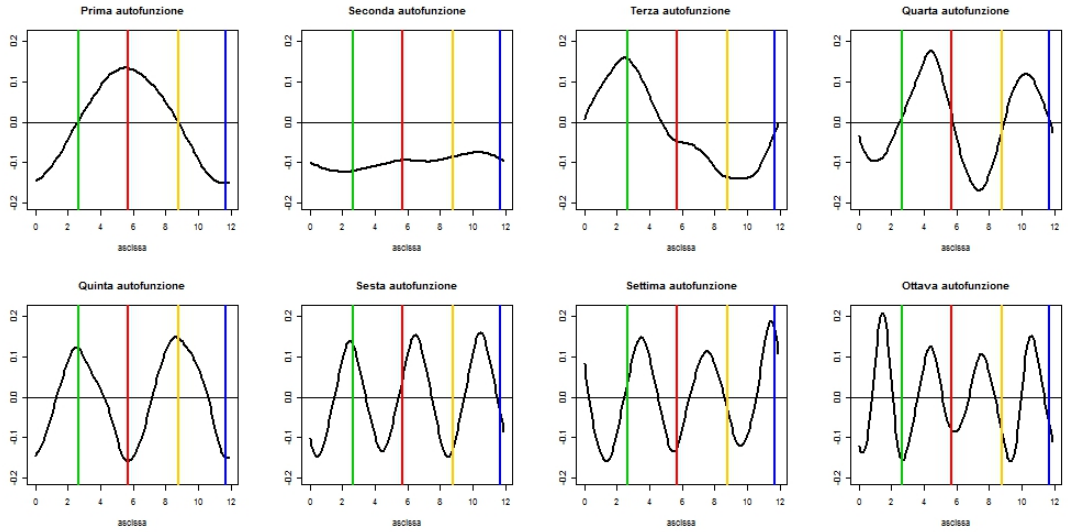


Figura 4.53: Prime otto autofunzioni ottenute tramite FPCA sul dataset composto dai dati di sintesi ottenuti dopo una tassellazione di Voronoi. Linea verde=equinozio di primavera (20 marzo), linea rossa=solstizio d'estate (21 giugno), linea arancione=equinozio d'autunno (22 settembre), linea blu=solstizio d'inverno (21 dicembre)

dei dati di sintesi nello spazio a tre dimensioni precedentemente ottenuto, utilizzando un metodo di clustering gerarchico. Scegliamo quindi come misura della distanza tra i dati, la distanza in L^2 tra le proiezioni dei dati di sintesi ottenute dalla FPCA, e per il linkage il metodo Ward.

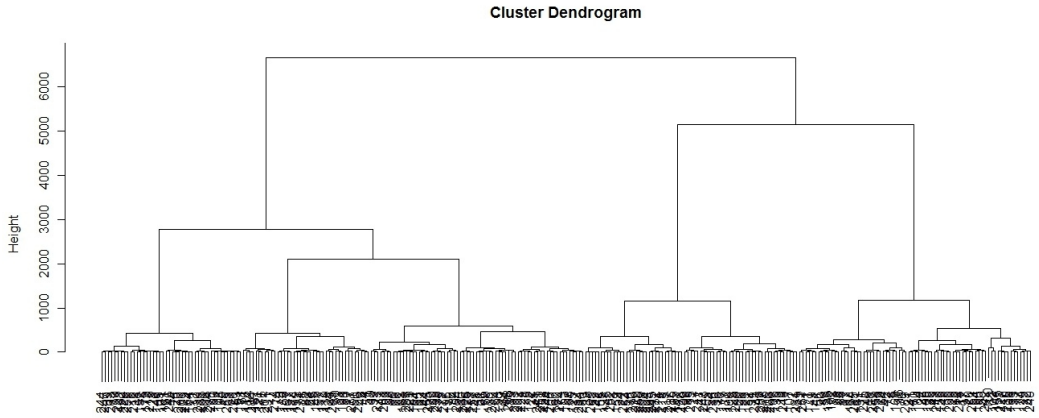


Figura 4.54: Dendrogramma relativo alla classificazione gerarchica delle proiezioni dei dati di sintesi di una tassellazione di Voronoi sullo spazio tridimensionale identificato dalle prime tre componenti principali, con distanza L^2, e linkage Ward

In Figura 4.54 presentiamo il dendrogramma relativo alla classificazione dei dati di sintesi di un'iterazione dell'algoritmo proposto. Osservando tale immagine, notiamo che il numero di cluster suggerito dal dendrogramma è $k = 3$ oppure $k = 5$. Ricordiamo che si tratta tuttavia di un risultato relativo ad una singola iterazione dell'algoritmo: è necessario quindi verificarne la stabilità lungo le diverse iterazioni, per esempio confrontando l'entropia associata alle diverse classificazioni finali ottenute.

4.5.3 Analisi dei risultati

Applichiamo, ora, l'algoritmo di classificazione spaziale descritto nel Capitolo 2, con le scelte precedentemente illustrate, ai dati funzionali ottenuti dalle osservazioni mensili dell'energia massima da immagazzinare in un sistema di backup.

Ricordiamo che, avendo scelto di effettuare la classificazione utilizzando un metodo gerarchico ad ogni iterazione, per problemi di ordine computazionale causati dalle grandi dimensioni del dataset, non è possibile effettuare una classificazione non spaziale dei dati di partenza (simple clustering).

Per scegliere il numero di cluster nei quali effettuare la suddivisione, procediamo come nel Paragrafo 4.3.4, ragionando in base all'entropia associata alle classificazioni ottenute. L'andamento dell'entropia media in funzione del numero di cluster utilizzato per la suddivisione è presentato in Figura 4.55. Dal grafico in Figura, notiamo che ci sono due minimi di entropia. Il minimo assoluto è raggiunto con $k = 5$ cluster, ed è coerente con quanto avevamo osservato a partire dal dendrogramma associato ad una singola classificazione (Figura 4.54). Procedendo nella suddivisione, si osserva un minimo relativo in corrispondenza di $k = 10$, tuttavia tale valore non ha un riscontro con quanto osservato dal dendrogramma; inoltre l'andamento dell'entropia è approssimativamente costante tra 8 e 11, di conseguenza il minimo osservato per $k = 10$ può non essere significativo. Infine, il valore dell'entropia media ottenuto per $k = 5$ è significativamente minore di quello ottenuto per $k = 10$.

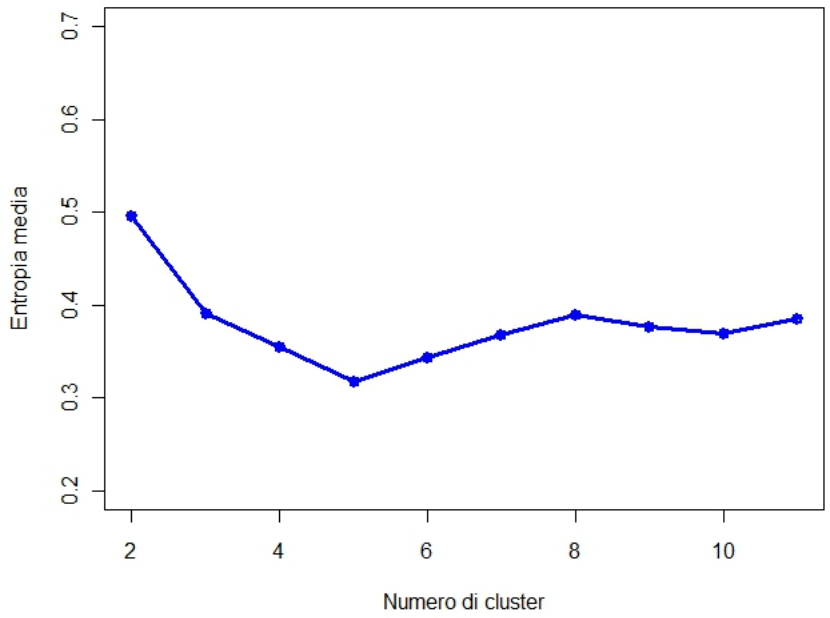
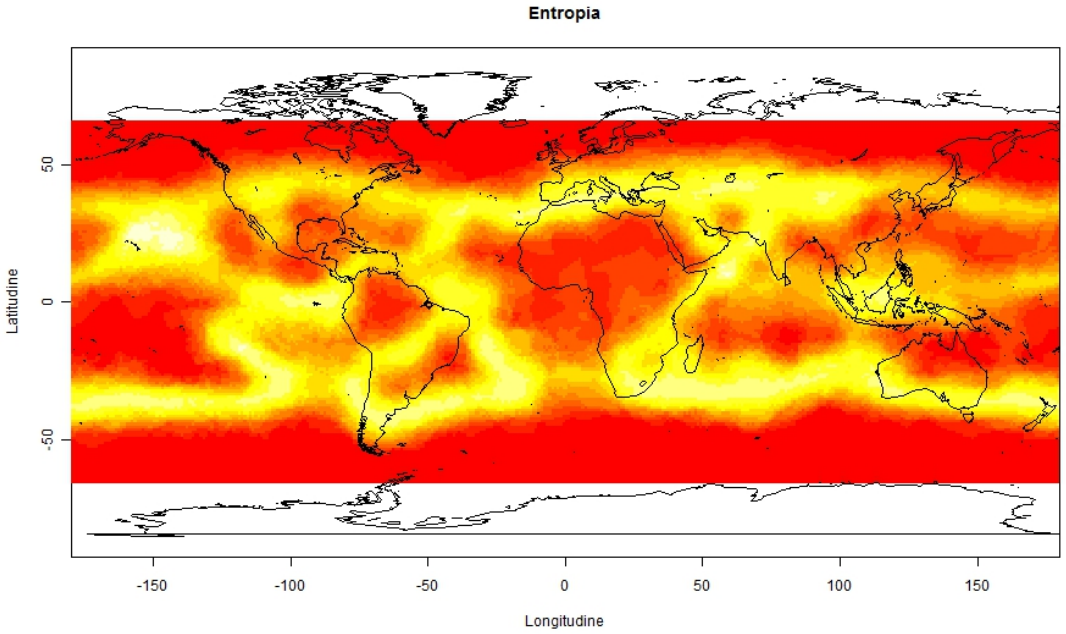
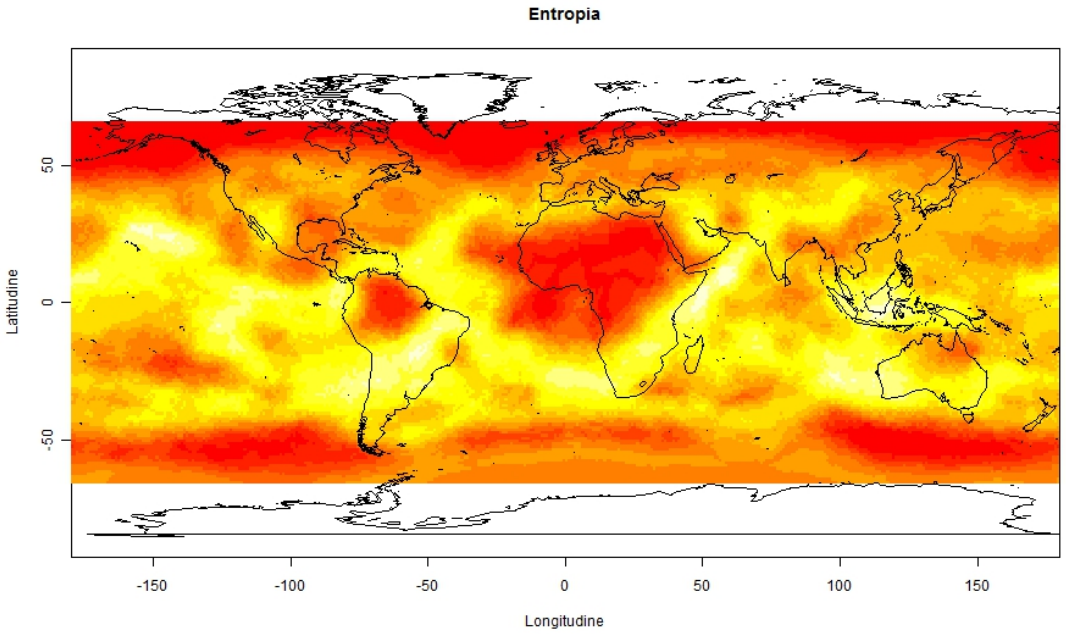


Figura 4.55: Andamento dell'entropia media all'aumentare del numero di cluster

In Figura 4.56, osserviamo infatti la mappa dell'entropia associata ad ogni sito nei due casi. Osserviamo che il risultato ottenuto per $k = 5$ è sensibilmente migliore di quello ottenuto per $k = 10$. Per le ragioni elencate, scegliamo la classificazione ottenuta per $k = 5$ cluster come migliore stima ottenuta per quanto riguarda la clusterizzazione spaziale del nuovo dataset analizzato.



(a) $k = 5$



(b) $k = 10$

Figura 4.56: Entropia della classificazione ottenuta con il metodo dello spatial clustering; i valori vicini a 1 sono mostrati in bianco, quelli vicini a 0 in rosso

In Figura 4.57, mostriamo la stima finale della classificazione ottenuta per k = 5. Mostriamo inoltre, grazie al plot dell'entropia associata alla classificazione in figura 4.54, che i risultati ottenuti lungo le diverse iterazioni sono molto stabili. Tuttavia, i dati funzionali associati a ciascuno dei cluster ottenuti sono difficili da visualizzare, a causa dell'elevata variabilità e dimensione del dataset di origine. Per analizzare i dati associati ai diversi cluster, quindi, mostriamo le curve associate ai cluster in una singola iterazione dell'algoritmo in Figura 4.58. In particolare, nei due pannelli della Figura presentiamo i dati di sintesi stessi (appartenenti, quindi, ad uno spazio infinito dimensionale) nel panello (a), e le proiezioni dei dati di sintesi sullo spazio tridimensionale identificato dalle prime tre componenti principali nel pannello (b): in entrambi i pannelli, i dati sono stati colorati in base alla classificazione k-medie effettuata nello spazio generato dalla prime 3 componenti principali.

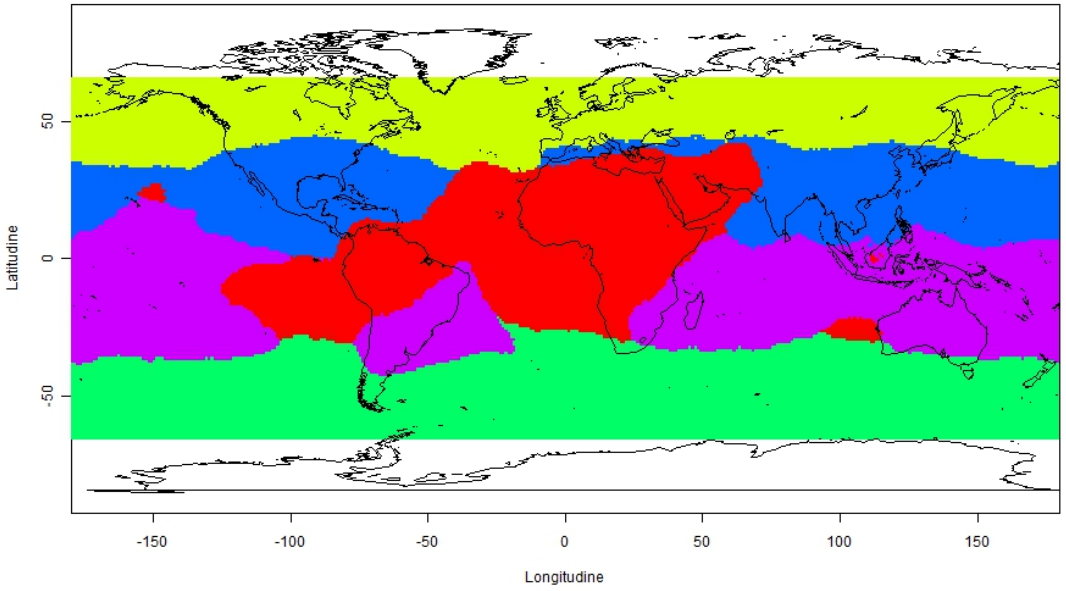
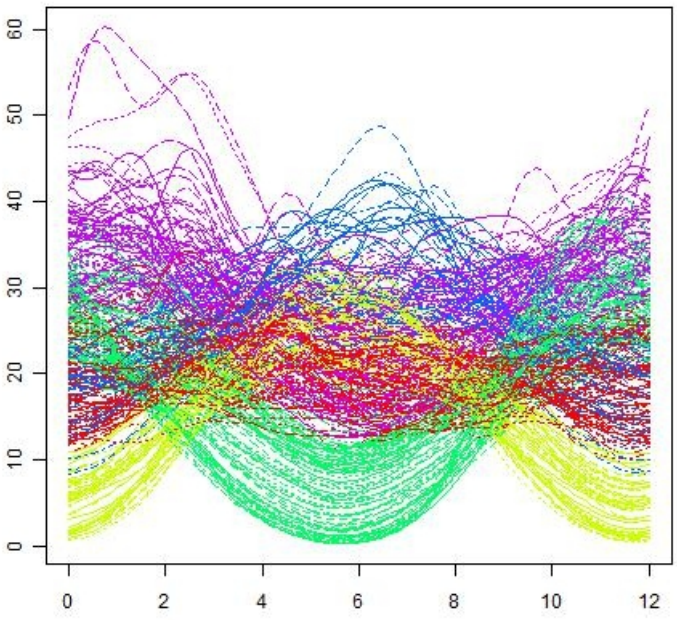


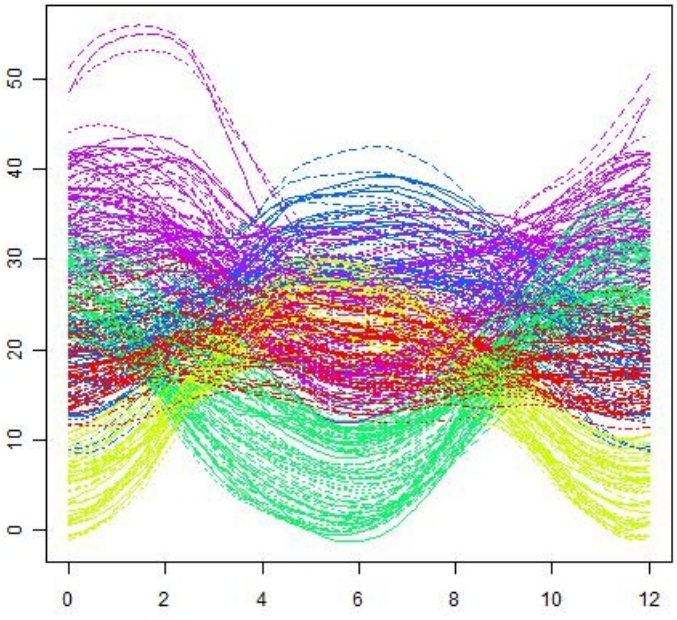
Figura 4.57: Stima finale della clusterizzazione ottenuta con la tecnica di classificazione spaziale proposta per $k = 5$

Confrontando i due pannelli in Figura 4.58, notiamo innanzitutto che le proiezioni dei dati di sintesi sullo spazio tridimensionale ottenuto tramite FPCA mantengono lo stesso andamento dei dati di sintesi infinito dimensionali, pur riducendone in maniera significativa il rumore. Notiamo anche che la classificazione ottenuta con il metodo proposto, sui dati tridimensionali, rispecchia il reale andamento dei dati di sintesi.

Analizzando quindi dal punto di vista climatico i risultati ottenuti, notiamo dalle Figure 4.57 e 4.58 come i cluster si possano spiegare nel modo seguente: otteniamo due cluster (verde e giallo) associati a zone subpolari, nelle quali l'insolazione è molto bassa e l'energia massima richiesta dal sistema di backup ha un andamento stagionale, opposto nei due cluster. Poi, otteniamo altri due cluster (blu e viola) suddivisi in base alla latitudine, nei quali l'insolazione media è maggiore e l'andamento annuale del parametro di riferimento è sempre stagionale. Questi due cluster, infatti, sono separati da quelli subpolari esclusivamente rispetto al valore medio annuale del parametro di riferimento. Infine, otteniamo un ultimo cluster (in rosso) che comprende il continente



(a) Dati di sintesi appartenenti a uno spazio funzionale infinito dimensionale



(b) Proiezioni dei dati di sintesi sullo spazio tridimensionale identificato dalle prime tre PC

Figura 4.58: Dati di sintesi ottenuti dopo una tassellazione di Voronoi, classificati tramite clustering gerarchico per $k = 5$, nello spazio infinito dimesionale (a) e tridimensionale ottenuto tramite FPCA (b)

africano, la penisola araba e l'America equatoriale, che corrisponde a dati che si situano mediamente all'interno del dataset, ma che non hanno tendenza stagionale. Le zone associate a questo cluster, quindi, corrispondono a regioni nelle quali l'insolazione è mediamente elevata durante tutto l'anno, mentre i periodi consecutivi senza sole sono molto brevi, e privi di una tendenza stagionale.

Osservando nuovamente la mappa presentata in Figura 4.57, notiamo poi che la classificazione non dipende esclusivamente dalla latitudine dei siti, ma che anche la longitudine influenza la disposizione spaziale dei cluster. Infatti, mentre nel continente americano sono presenti zone associate a tutti e cinque i cluster in proporzioni tra loro simili, e la transizione tra le diverse zone è graduale, in corrispondenza delle longitudini comprendenti Africa ed Europa, invece, c'è una predominanza del cluster associato a dati non stagionali, con una transizione molto veloce verso zone associate a cluster più stagionali. Se osserviamo l'Europa meridionale, infatti, vediamo che tale zona risulta di fatto suddivisa in tre cluster: uno che ricopre la parte continentale, uno in corrispondenza del bacino del mediterraneo, e un ultimo che comprende la Sicilia e la Grecia; in effetti, in questa zona del mondo il clima varia molto rapidamente rispetto alla latitudine e alle caratteristiche geografiche, e i microclimi che si riescono ad osservare all'interno di tale zona, pur di dimensioni relativamente ristrette, sono molti: dal clima alpino presente in Svizzera, a quello mediterraneo dell'Italia centrale, a quello molto secco quasi desertico dell'Italia meridionale e della Grecia. Infine, in Asia il cluster corrispondente a zone nelle quali l'andamento del parametro è non stagionale non è presente; le zone del mondo appartenenti al continente asiatico, dunque, vengono suddivise nei soli quattro cluster corrispondenti ad andamenti stagionali della variabile considerata, e quasi esclusivamente secondo la latitudine.

Conclusioni e sviluppi futuri

L'obiettivo posto all'inizio di questo lavoro di tesi era quello di sviluppare ed analizzare delle tecniche per risolvere un problema di classificazione non supervisionata di dati funzionali in ambito spaziale.

Al fine di risolvere tale problema, abbiamo in un primo momento introdotto e formalizzato modelli adatti a spiegare la dipendenza spaziale nei dati. In particolare, sono stati descritti ed analizzati due modelli di dipendenza spaziale, cioè variogramma e Hidden Random Markov Field, che ci sono successivamente stati utili per gli studi di simulazione, nel momento in cui è stato necessario generare dati spazialmente dipendenti.

Successivamente, abbiamo introdotto un algoritmo generale e flessibile per classificare dati funzionali spazialmente referenziati che tiene opportunamente conto della struttura di dipendenza spaziale esistente, e che agisce su rappresentanti locali dei dati ottenuti tramite la creazione di una tassellazione di Voronoi dello spazio considerato.

Tale algoritmo è stato poi analizzato per mezzo di opportune simulazioni, per esplorarne le prestazioni al variare dei parametri del modello di generazione dei dati, e di alcune scelte nella definizione dell'algoritmo stesso. Per quanto riguarda il modello di generazione dei dati, in una prima fase del lavoro abbiamo supposto di osservare la realizzazione di un Hidden Random Markov Field, ovvero abbiamo ipotizzato che i dati fossero realizzazioni di funzioni aleatorie indipendenti condizionatamente alle etichette del campo latente. In un secondo momento, abbiamo sviluppato un modello misto, nel quale la media delle osservazioni dipende dalle etichette di un Markov Random Field latente, ma non vale l'ipotesi di indipendenza condizionata rispetto al campo latente. In particolare, abbiamo introdotto una covarianza tra le osservazioni nei diversi siti, utilizzando un noto modello di covariogramma. Gli studi di simulazione hanno permesso di trarre due conclusioni: innanzitutto, l'approccio presentato permette di ottenere delle classificazioni migliori rispetto alle tecniche classiche esistenti in ambito non spaziale; in secondo luogo, i risultati ottenuti sono migliori se i dati provengono effettivamente da un Hidden Random Markov Field, cioè se è rispettata l'ipotesi di indipendenza delle osservazioni condizionatamente alle etichette.

In particolare, abbiamo studiato le prestazioni del metodo proposto al variare del numero di tasselli del diagramma di Voronoi, mostrando empiricamente l'esistenza di un numero ottimo di tasselli in corrispondenza del quale la classificazione ottenuta è migliore, in termini di errore di misclassificazione delle etichette di partenza. A questo proposito, abbiamo studiato il comportamento di tale punto di ottimo, evidenziando il fatto che non si tratta di un valore costante, ma dipende dal modello con il quale i dati sono stati generati; in particolare osserviamo una dipendenza dalla distanza tra le densità associate alle distribuzioni che generano il segnale condizionatamente al valore assunto dall'etichetta, dalla variabilità associata ai dati e dal tipo di modello utilizzato per descrivere la dipendenza spaziale.

Abbiamo infine preso in considerazione un caso reale, analizzando un dataset che si situa nel contesto proposto, cioè quello di dati funzionali con dipendenza spaziale osservati su un reticolo regolare. I dati presi in esame riguardano infatti l'evoluzione nel tempo di un parametro climatico (il numero equivalente di giorni consecutivi senza sole), osservato su un reticolo formato da meridiani e paralleli che ricopre l'intera superficie terrestre, ad eccezione delle calotte polari.

I problemi da risolvere per effettuare l'analisi del dataset proposto sono stati molteplici. Innanzi tutto, si è posto il problema di passare, tramite tecniche classiche di smoothing introdotte nell'ambito dell'analisi funzionale, dalle 12 osservazioni discrete dell'evoluzione del parametro in questione nel tempo per ogni sito, a un dato funzionale periodico e vincolato, valutato su una griglia fine di ascissa.

Una volta creato il set di dati funzionali, è stato necessario declinare la tecnica generale di classificazione precedentemente descritta ed analizzata al caso reale in questione; in particolare, le modifiche hanno riguardato il fatto che il reticolo regolare in corrispondenza del quale i dati sono osservati non si trovasse su un piano, ma su una superficie sferica come quella terrestre. Abbiamo, dunque, scelto di utilizzare la distanza geodetica per creare il diagramma di Voronoi e i dati di sintesi di ogni tassello, la FPCA per ridurre la dimensione dei dati di sintesi ad ogni iterazione e il metodo delle k–means per la classificazione dei vettori q–dimensionali degli scores. Successivamente, è stata effettuata l'intera analisi e classificazione del set di dati in questione con l'utilizzo dello schema proposto.

Un altro problema affrontato nel corso dell'analisi è stato la declinazione di un metodo per la scelta del numero di cluster. Infatti, trattandosi di un problema di classificazione non supervisionata, il numero di classi in cui dividere il dataset non è noto a priori. Per effettuare tale scelta, è stata introdotta ed utilizzata l'entropia associata alla classificazione, un indice quantitativo associato alla nitidezza delle immagini finali ottenute.

Infine, abbiamo effettuato degli studi di robustezza dei risultati ottenuti, variando in un primo momento il numero q di componenti principali da considerare per la riduzione dimensionale dei dati di sintesi, ed in un secondo momento la tecnica stessa utilizzata per la classificazione dei vettori degli scores, classificando tali vettori ad ogni passo per mezzo di un metodo di clusterizzazione gerarchica.

Alla luce delle analisi effettuate abbiamo concluso che il metodo di classificazione spaziale proposto ha delle prestazioni migliori rispetto ai metodi classici in termini di definizione e nitidezza delle mappe risultanti. La classificazione ottenuta non sembra associabile ad una mistura di distribuzioni, quanto piuttosto ad una segmentazione del dataset; esso cioè viene suddiviso a livello spaziale, in aree significative, all'interno delle quali i dati sono omogenei, al fine di dare una rappresentazione più nitida del fenomeno in questione. Infine, concludiamo che i risultati ottenuti, nonché la scelta del numero ottimale di cluster, sono molto robusti al variare delle tecniche utilizzate per effettuare l'analisi.

Abbiamo visto, nel corso di questo lavoro, che la tecnica di classificazione proposta ha delle ottime prestazioni, confrontata con le tecniche classiche esistenti. Questo, tuttavia, non conclude le ricerche da effettuare nell'ambito della classificazione di dati spazialmente dipendenti; al contrario apre nuove possibili vie per la ricerca in questo campo.

Osserviamo, infatti, che il risultato finale dell'algoritmo, che è iterativo, è ugualmente influenzato dai risultati parziali ottenuti in ciascuna iterazione. Tuttavia, alcune classificazioni sono localmente migliori di altre, nel senso che dovremmo preferire quelle

classificazioni ottenute a partire da tassellazioni di Voronoi che si adattano meglio ai confini tra i diversi cluster. Un possibile approccio per tenere conto di questa considerazione, dal momento che i confini tra i cluster sono incogniti, è associare, ad ogni tassello ottenuto dal diagramma di Voronoi, un peso inversamente proporzionale rispetto alla varianza dei dati all'interno del tassello stesso, e tenere in considerazione per ogni sito i pesi ottenuti nelle diverse iterazioni per effettuare la stima finale. In questo modo, il risultato ottenuto nel singolo sito dalla classificazione di un tassello che contiene dati molto variabili e dunque probabilmente associati a differenti distribuzioni del segnale, verrebbe pesato meno rispetto al risultato ottenuto nello stesso pixel dalla classificazione di un tassello che contiene dati con variabilità molto bassa.

Un altro problema, di natura più teorica, che abbiamo riscontrato nel corso delle analisi fatte è il seguente: utilizzando la FPCA per ridurre dimensionalmente il dataset di sintesi ad ogni iterazione del metodo, stiamo di fatto cambiando lo spazio di rappresentazione dei dati. Il problema sorge dal fatto che, al fine di ottenere una stima finale della classificazione, confrontiamo risultati ottenuti in iterazioni diverse, senza tenere in considerazione il fatto che la base sulla quale vengono proiettati i dati cambia. La soluzione pratica che viene adottata è la seguente: partendo dall'ipotesi che le variazioni della base siano piccole, effettuiamo un matching tra le componenti ottenute in iterazioni diverse, per identificare e mantenere le componenti più simili a quelle scelte alla prima iterazione. Lo svantaggio associato a questa tecnica è che non ci sono risultati teorici che assicurano che le variazioni della base tra un'iterazione e l'altra siano, effettivamente, piccole. Una soluzione teorica al problema consisterebbe nell'utilizzo di una base fissata per proiettare i dati, come per esempio la base di Fourier. Tuttavia, tale approccio ha lo svantaggio di utilizzare una base non più data–driven, e rimarrebbe il problema di dover scegliere a priori una base conveniente. Una possibile soluzione, che necessita un ulteriore sviluppo, è quella di trovare, a partire dai dati, un'unica base ortonormale ottima. Tale base può essere definita come la media delle autofunzioni trovate nelle diverse iterazioni dell'algoritmo, vincolata in modo da mantenere la proprietà di ortonormalità della base.

Bibliografia

[Banerjee et al., 2004] Banerjee, S., Carlin, B. and Gelfand, A. (2004). Hierarchical modeling and analysis for spatial data. Chapman & Hall.

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) *36*, 192–236.

[Cressie, 1991] Cressie, N. (1991). Statistics for Spatial Data (Wiley Series in Probability and Statistics). John Wiley & Sons. New York.

[Dubes et al., 1990] Dubes, R., Jain, A., Nadabar, S. and Chen, C. (1990). MRF model-based algorithms for image segmentation. In Pattern Recognition, 1990. Proceedings., 10th International Conference on vol. 1, pp. 808–814, IEEE.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis. Springer Verlag.

[Geman and Geman, 1993] Geman, S. and Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. Journal of Applied Statistics *20*, 25–62.

[Geman and Graffigne, 1986] Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. In Proceedings of the International Congress of Mathematicians vol. 1496, p. 1517,.

[Härdle, 1991] Härdle, W. (1991). Smoothing techniques: with implementation in S. Springer.

[Johnson and Wichern, 2002] Johnson, R. and Wichern, D. (2002). Applied multivariate statistical data analysis. Prentice Hall: Upper Saddle River, NJ.

[Ke and Wang, 2001] Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. Journal of the American Statistical Association *96*, 1272–1298.

[Künsch et al., 1995] Künsch, H., Geman, S. and Kehagias, A. (1995). Hidden Markov Random Fields. The annals of applied probability *5*, 577–602.

[Leung and Lam, 2002] Leung, C. and Lam, F. (2002). Maximum a posteriori spatial probability segmentation. In Vision, Image and Signal Processing, IEE Proceedings-vol. 144, pp. 161–167, IET.

[NASA, 2008] NASA (2008). Surface meteorology and Solar Energy. http://eosweb.larc.nasa.gov.

[Ramsay and Silverman, 2002] Ramsay, J. and Silverman, B. (2002). Applied functional data analysis: methods and case studies. Springer Verlag.

[Ramsay and Silverman, 2005] Ramsay, J. and Silverman, B. (2005). Functional data analysis. 2 edition, Springer Verlag.

[Sangalli et al., 2009] Sangalli, L., Secchi, P., Vantini, S. and Veneziani, A. (2009). A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. Journal of the American Statistical Association *104*, 37–48.

[Secchi et al., 2011] Secchi, P., Vantini, S. and Vitelli, V. (2011). A clustering algorithm for spatially dependent functional data. Technical report MOX.